

# **STATISTICAL METHODS FOR GENOTYPE ASSAY DATA**

by

**Soo Yeon Cheong**

BSc in Information Statistics, Hankuk University of Foreign Studies, South Korea, 2000

MSc in Statistics, Seoul National University, South Korea, 2003

Submitted to the Graduate Faculty of  
the Graduate School of Public Health in partial fulfillment  
of the requirements for the degree of  
**Doctor of Philosophy**

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH  
GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

By

**Soo Yeon Cheong**

It was defended on

**April 13, 2010**

and approved by

Dissertation Advisor:  
Eleanor Feingold, PhD  
Professor  
Human Genetics  
Graduate School of Public Health  
University of Pittsburgh

Committee Member:  
Chien-Chen (George) Tseng, ScD  
Assistant Professor  
Biostatistics  
Graduate School of Public Health  
University of Pittsburgh

Committee Member:  
Yan Lin, PhD  
Research Assistant Professor  
Biostatistics  
Graduate School of Public Health  
University of Pittsburgh

Committee Member:  
M. Michael Barmada, PhD  
Associate Professor  
Human Genetics  
Graduate School of Public Health  
University of Pittsburgh

Copyright © by Soo Yeon Cheong

2010

# STATISTICAL METHODS FOR GENOTYPING ASSAY DATA

Soo Yeon Cheong, Ph.D.

University of Pittsburgh, 2010

There are many methods to detect any relationship between genotype and phenotype. All of them need to be preceded by measuring genotypes. Genotypes are assigned at each marker for every person to be tested based on raw data from any of a number of different assays. After genotyping, association is tested with a  $\chi^2$  test on a  $2 \times 3$  table of phenotype  $\times$  genotype for a simple case-control study design. Based on the  $\chi^2$  test, we may infer that one of the alleles at the marker might increase risk of the disease. In this dissertation we study analysis methods for raw data from genotyping assays, with particular attention to two issues: genotype calling for trisomic individuals, and design and testing for pooled DNA studies.

There are a number of statistical clustering techniques and software packages in use to call genotypes for disomic individuals. However, standard software packages cannot be used if a chromosomal abnormality exists. We used data from individuals with Down syndrome, who have an extra copy of chromosome 21. A method of calling genotypes for individuals with Down syndrome was already suggested in a previous study. In this study we propose a new method to improve the genotype calling in this situation.

In most association studies, individual genotyping is used, but that approach has high cost. Pooled genotyping is a cost effective way to perform the first stage of a genetic association study. DNA pools are formed by mixing DNA samples from multiple individuals before genotyping. Pooled DNA is assayed on a standard genotyping chip, and allele frequencies are estimated from

the raw intensity data for the chip. Many previous studies looked at the issue of estimating more accurate allele frequencies for pooled genotyping. In this study we consider two different issues: design of pooled studies and statistical testing methods. We consider several pooling designs with the same cost and compare to figure out the most effective design. And we also discuss the most appropriate statistics for testing each design.

The two issues addressed in this study are pre-requisites to any genetic association analysis. Genetic association studies are leading to new knowledge that will eventually improve prevention and treatment options for many diseases. However, these studies cannot succeed unless we know how to design and analyze them correctly. Using incorrect genotype calls, incorrect statistics, or inefficient designs will all severely compromise the public health advances that these studies are able to make. The studies we have done will help lead to more correct and efficient genetic association studies, and thus to quicker and surer advances in prevention and treatment. Thus this work has great public health significance.

## TABLE OF CONTENTS

<b>PREFACE.....</b>	<b>XII</b>
<b>1.0 INTRODUCTION .....</b>	<b>1</b>
<b>1.1 GENETIC ASSOCIATION STUDIES .....</b>	<b>1</b>
<b>1.2 GENOTYPING TECHNOLOGIES.....</b>	<b>2</b>
<b>1.3 DATA FROM GENOTYPING METHODS .....</b>	<b>3</b>
<b>1.3.1 Raw Data .....</b>	<b>3</b>
<b>1.3.2 Data from Pooled DNA.....</b>	<b>4</b>
<b>1.3.3 Genotype Calling.....</b>	<b>5</b>
<b>1.4 OVERVIEW AND PROBLEMS WE CONSIDERED.....</b>	<b>7</b>
<b>1.4.1 Genetic Association Studies with Pooled Genotyping .....</b>	<b>7</b>
<b>1.4.2 Genotype Calling Methods for Trisomic Samples .....</b>	<b>11</b>
<b>1.5 PURPOSE OF THIS DISSERTATION.....</b>	<b>14</b>
<b>1.6 DATASETS.....</b>	<b>16</b>
<b>1.6.1 Pancreatic Cancer Pooled Data and Case-Control Study.....</b>	<b>16</b>
<b>1.6.2 Down Syndrome Data for Genotype Calling.....</b>	<b>16</b>
<b>2.0 GENOTYPE CALLING FOR TRISOMIC SAMPLES .....</b>	<b>18</b>
<b>2.1 BACKGROUND .....</b>	<b>18</b>
<b>2.2 PREVIOUS GENOTYPE CALLING METHODS .....</b>	<b>21</b>

2.3	OUR PROPOSED METHOD FOR TRISOMIC GENOTYPE CALLING.....	23
2.4	EXAMPLE DATASETS.....	26
2.5	COMPARISON OF METHODS.....	26
2.6	RESULTS .....	29
2.6.1	Genotype Calling by HINM.....	29
2.6.2	Comparison of Results of HINM with Lin's IBM .....	30
2.6.2.1	Apply both HINM and Lin's IBM methods.....	30
2.6.2.2	Count Mismatches.....	32
2.6.2.3	Examples for Visual Detection .....	35
2.7	CONCLUSION.....	40
3.0	MODELS, TEST STATISTICS, AND DESIGNS FOR GENETIC ASSOCIATION	
	STUDIES WITH POOLED GENOTYPING .....	42
3.1	INTRODUCTION.....	42
3.2	MODEL OF POOLING VARIABILITY .....	45
3.3	DESIGNS .....	47
3.3.1	Three Designs for a Case-Control Study .....	47
3.3.2	Allele Frequency Estimates from Each Design .....	48
3.4	TEST STATISTICS.....	49
3.5	PANCREATIC CANCER AFFYMETRIX 6.0 POOLED DATA.....	51
3.6	RESULTS .....	52
3.6.1	Relative Efficiency of the Three Designs .....	52
3.6.2	Comparison of the Three Test Statistics using the Affymetrix 6.0 Pooled Data .....	53

3.6.3	What is the Best Pooling Design? .....	56
3.6.4	What is the most Appropriate Test Statistic? .....	57
3.6.5	Designs with Covariates .....	58
4.0	CONCLUSION AND DISCUSSION .....	60
4.1	NEW GENOTYPE CALLING METHOD FOR TRISOMIC INDIVIDUALS...	60
4.2	MOST EFFICIENT POOLING STRATEGES AND TEST STATISTICS.....	62
4.3	MORE GENERAL COMMENTS ON USES FOR RAW GENOTYPE DATA AND RELATED PROBLEMS .....	63
APPENDIX A. MEAN AND VARIANCE OF THE ALLELE FREQUENCY .....		67
APPENDIX B. VARIANCE OF THE ALLELE FREQUENCY FOR TEST STATISTICS.. .....		69
B.1.	Variance of the Allele Frequency for Three Designs.....	69
B.2.	Predicted Variance of the Overall Allele Frequency.....	72
BIBLIOGRAPHY .....		73



## LIST OF TABLES

Table 2.1: Number (Percentage) of SNPs of each quality in each dataset .....	31
Table 2.2: Percentage of SNPs that could be called with each algorithm .....	31
Table 2.3: Number of randomly selected 200 SNPs worked for both methods .....	32
Table 2.4: Number of mismatch SNPs by mismatch rate (MMR) and SNP quality .....	33
Table 2.5: Number of SNPs of quality and/or mismatch rate (MMR) by categorized call results for both Dataset 1 and Dataset 2 .....	34
Table 3.1: True allele frequencies of three designs .....	49
Table 3.2: Allele frequency table for $\chi^2$ -test.....	51
Table 3.3: Number of significant SNPs among selected 200 SNPs of modified t-test and $\chi^2$ -test at $\alpha = 0.05$ .....	55

## LIST OF FIGURES

Figure 1.1: Scatter plots for one SNP .....	6
Figure 1.2: Wrong clusters by Lin's IBM for trisomic individuals.....	13
Figure 2.1: Examples of genotype calling results of various SNPs using K-means.....	20
Figure 2.2: Example of Genotype Clusters Incorrectly Found by Genotype Calling Method .....	21
Figure 2.3: Skewness of Homozygote Clusters .....	22
Figure 2.4: Examples of genotype calling results for trisomic individuals with misclassified clusters by Lin's IBM .....	23
Figure 2.5: Example of "Good" and "Bad" calls of high and low quality SNPs .....	29
Figure 2.6: Examples of genotype calls by HINM .....	30
Figure 2.7: Good call results of both Lin's IBM and HINM.....	36
Figure 2.8: Genotype calls for SNP with low-intensity curvature by Lin's IBM and HINM.....	37
Figure 2.9: Bad genotype calls for both Lin's IBM and HINM.....	38
Figure 2.10: Better genotype calls by Lin's IBM than HINM .....	39
Figure 2.11: Better genotype calls by HINM than Lin's IBM .....	40
Figure 3.1: True allele frequencies of population, samples, and pool .....	45
Figure 3.2: Three designs for non-covariate model .....	48
Figure 3.3: Pooled affymetrix 6.0 pancreatic cancer data .....	52
Figure 3.4: Three designs of covariate model.....	58

Figure 4.1: Generalized genotyping.....	65
---	----

## **PREFACE**

First I would like to express gratitude to my advisor Professor Eleanor Feingold for her suggestions, patience, and constant support. She introduced me to this exciting project and fully supported me through my research. Her excellent guidance with patience always led me in the right direction, and let me have completed my thesis. I am honored to be her student.

I also would like to thank my other committee members, Professor George Tseng, Professor Michael M. Barmada, and Professor Yan Lin for their helpful suggestions and encouragement.

And I wish to express my appreciation to the rest of faculty, staffs, and students of the Department of Biostatistics and Human Genetics at University of Pittsburgh for the education and generous support during the past few years. And I also would like to thank the students, postdocs, and faculties of the statistical genetics group. The seminars organized by this group are very interesting and helpful to me.

I would like to thank my family and friends for being always with me and for encouraging me all the time.

## 1.0 INTRODUCTION

### 1.1 GENETIC ASSOCIATION STUDIES

Genetic association studies look for correlation between genotype and phenotype. There are many methods for doing that, depending on study design, but all of them require that we start by measuring genotypes. Each person in the study must be assigned a genotype at each marker that is to be tested. Typically these are SNP markers with two alleles. For a simple case-control study design, association is then tested with some kind of  $\chi^2$  test on the  $2 \times 3$  table of phenotype  $\times$  genotype. Based on the  $\chi^2$  test, a difference in the allele frequencies or genotype frequencies between two cohort groups tests whether the genetic marker might be associated with risk of the disease. One variation on this is a “pooled” association study, in which people are not genotyped individually. Rather, DNA from study subjects is combined and the genotyping assay yields not individual genotypes but group estimates of allele frequencies (discussed further below). A number of different assays are available for genotyping people individually or in pools, including several high-throughput “chips.” These assays produce several levels of raw data, which are then processed into genotype data. In most standard uses of genotyping chips, only the called genotypes are used. Genotypes are used for linkage analysis as well, but that raw data is even less often used in linkage studies than in association studies. The raw intensity data is generally ignored. But the raw data is important for a number of special problems, some of which are

addressed in this dissertation. The raw data is often also used for studies of copy number variation, but methods for copy number studies will not be a part of this dissertation. Specifically, in this dissertation I will look at two problems involving raw data from genotyping chips: how to design and analyze pooled DNA association studies, and how to call genotypes for trisomic data.

## 1.2 GENOTYPING TECHNOLOGIES

There are several tens of companies that offer genotyping technologies, such as Agilent, Affymetrix, Perlegen, Illumina, and so on (Hardiman 2004; Perkel 2008). Among them, two gene chip platforms are the most popularly used: Affymetrix gene chip and Illumina gene chip. These chips are used to determine genotypes of each person at each marker in order to conduct genetic studies. And they are generally used for large-scale studies of up to hundreds of thousands of SNPs. However they have fairly different characteristics, though their application and formats are somewhat similar.

Affymetrix and Illumina gene chips are very different in many ways (Barnes et al., 2005; Maouche et al., 2008). Affymetrix produces gene chips by spotting oligonucleotides using photochemical *in situ* synthesis, while Illumina produces oligonucleotide bead-based arrays using standard oligonucleotide synthesis. Affymetrix gene chips look like a checkerboard and each probe is synthesized at a specific location, and Illumina gene chips use a decoding process to identify each probe's location on the array using genes' molecular addresses. Affymetrix uses multiple matching probes for each gene complemented by one-base mismatch probes which are controls for non-specific hybridization, while Illumina uses a random self-assembly process to put oligonucleotides on the array. Therefore there is no mismatch control for Illumina gene chips.

In addition, each Illumina array contains multiple samples unlike Affymetrix arrays (Barnes et al., 2005).

## 1.3 DATA FROM GENOTYPING METHODS

### 1.3.1 Raw Data

The raw data can be observed as an image file from scanning a chip. We can get intensities from decoding the brightness of each spot on the image file. Define intensities of allele  $A$  and allele  $B$  as  $y_A$  and  $y_B$ . Individuals with genotype  $AA$  would have a high value of  $y_A$  but a low value of  $y_B$ , while individuals with genotype  $BB$  would have low  $y_A$  and high  $y_B$ . And individuals with genotype  $AB$  would have similar values of  $y_A$  and  $y_B$  (Lin et al., 2008).

Usually there are two types of probes in Affymetrix chips: perfect match probes (PM) and mismatch probes (MM). PM is a completely complementary probe to the target sequence, while MM is complementary except for a single mismatched base. There are several algorithms to summarize probe intensities numerically. The algorithm using both PM and MM probes is the common method, which is to subtract of MM probe intensity from the PM probe intensity. However the PM-only algorithm is popular, which is to eliminate background noises from PM probe intensity. The purpose of these methods is to find the true probe intensity by removing noise from the PM probe intensity. For most Affymetrix chip data, subtraction of MM probe intensity is an accurate method to remove background (Dalma-Weiszhausz et al., 2006). Based on the intensities from these algorithms, genotypes can be called. Relative allele signal (RAS) scores are obtained by combining these intensity values using Affymetrix GeneChip DNA

analysis software (GDAS). RAS scores are defined as the ratio of the signal of the *A* allele to the sum of *A* and *B* alleles, that is  $RAS = y_A/(y_A+y_B)$ . RAS score should be close to 1 for *AA* genotypes, 0.5 for *AB* genotypes, and 0 for *BB* genotypes. Two different RAS scores can be measured from the same allele:  $RAS_1$  for the sense strand and  $RAS_2$  for the antisense strand. Thus RAS scores can be used to classify the genotypes, and also used for estimating the allele frequency in pooled DNA data (Kirov et al., 2006; Norton et al., 2002; Affymetrix Manual). On the other hand, intensities of Illumina chips are detected by decoding the array. Illumina chips give more or less directly intensities  $y_A$  and  $y_B$  for each allele. Several labeling channels are used to estimate – Cy3 (labeled red), Cy5 (labeled green), and no label. There are three labeling results after scanning the chips - red, green, or blank. The intensities from the channels (Cy3 and Cy5) represent two alleles of each SNP respectively. Genotypes of each SNP can be called by the genotyping software of Illumina, called GeneCall, based on the intensities (Fan et al., 2006).

### **1.3.2 Data from Pooled DNA**

Individual genotyping is the most popularly used and powerful method to examine the association between genetic factors and diseases in many studies (Bader et al., 2001). But if large numbers of individuals are genotyped, high cost is required for individual genotyping. However pooled genotyping can be more cost-effective than individual genotyping, and can be used as a prescreening method at the first stage in a genome-wide association study. DNA pools have to be constructed from DNA samples of multiple individuals before genotyping. DNAs from several people are mixed together in equal quantities and the mixture is assayed on a chip. The measured intensities can be interpreted as allele frequencies in the pool. However DNA pooling has not been used much because of concerns about bias and variability of pooled allele frequency



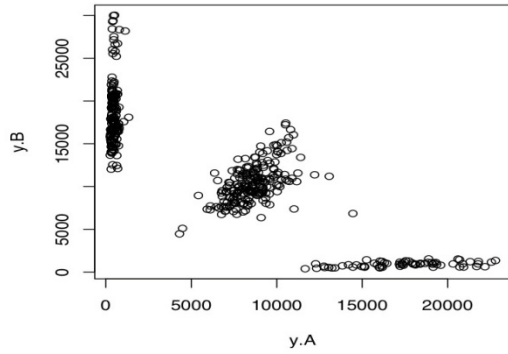
estimates. The bias is due to unequal hybridization of the two alleles and the variance is due to poor measurement of the “equal” quantities of DNA from each person as well as noise in reading the intensity off the chip.

### 1.3.3 Genotype Calling

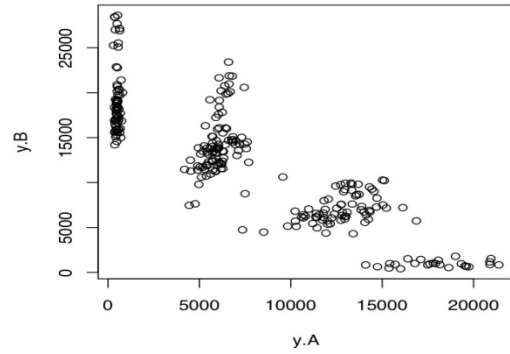
Disomic individuals have three possible genotypes, which can be thought of as  $AA$ ,  $AB$ , and  $BB$ . However trisomic individuals have four possible genotypes, which are  $AAA$ ,  $AAB$ ,  $ABB$ , and  $BBB$ . Figure 1.1 shows examples of raw intensity data and genotype calls for one SNP with lots of disomic or trisomic individuals, respectively. Each point in each plot represents an individual. Figure 1.1-(a) and (b) are example plots of the intensities  $y_A$  (x-axis) vs.  $y_B$  (y-axis). They show us three distinct genotype groups for disomic individuals and four groups for trisomic individuals. In Figure 1.1-(a), the cluster close to the x-axis represents  $AA$  genotypes, while the opposite one represents  $BB$  genotypes. The middle cluster between the  $AA$ - and  $BB$ - genotype groups represents  $AB$  genotypes. Similarly, Figure 1.1-(b) shows that the clusters close to the x- and y-axes represent  $AAA$  and  $BBB$  genotypes, respectively. The middle clusters, which are between the  $AAA$  and  $BBB$  groups, represent  $AAB$  and  $ABB$  genotypes. Therefore we could find and classify the genotypes of each SNP based on intensities of each allele using statistical clustering methods, then finally call the genotypes of the SNP of each individual. In most cases, the clustering is done on 1-dimensional data after transformation. Figure 1.1-(c) and (d) are the transformed plots of Figure 1.1-(a) and (b), using the formulas  $(y_A + y_B)$  as the x-axis and  $y_A / (y_A + y_B)$  as the y-axis. After the transformation, the y-axis value can be used for 1-dimensional clustering in order to call the genotypes. The range of the y-axis in Figure 1.1-(c) and (d) is 0 to 1. Then Figure 1.1-(c) shows that  $AA$  genotype group is close to the bottom (close to 0 in y-axis)

and *BB* genotype group is close to the top (close to 1 in y-axis). The middle cluster is for the *AB* genotype group. And Figure 1.1-(d) shows that the bottom cluster is for *AAA* genotypes and the top is for *BBB* genotypes. The second and third clusters from the bottom are for the *AAB* and *ABB* genotype groups, respectively.

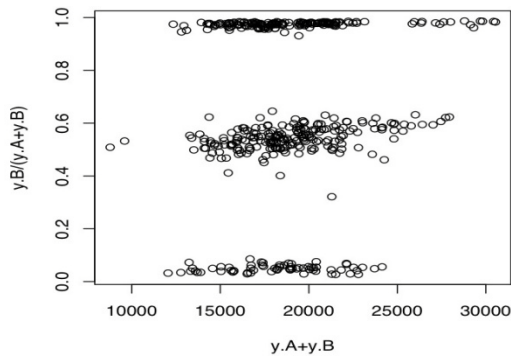
(a) Raw Intensity – Disomic



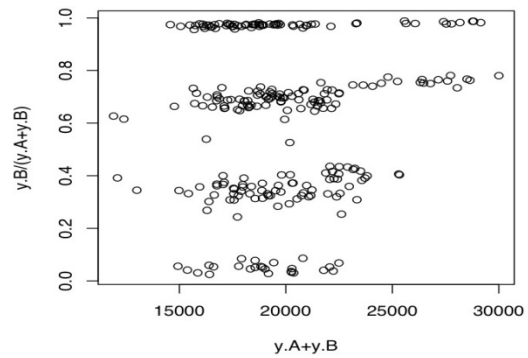
(b) Raw Intensity - Trisomic



(c) Transformed Intensity – Disomic



(d) Transformed Intensity – Trisomic



**Figure 1.1: Scatter plots for one SNP**

## **1.4 OVERVIEW AND PROBLEMS WE CONSIDERED**

### **1.4.1 Genetic Association Studies with Pooled Genotyping**

DNA pools are created by mixing DNA samples from multiple individuals before genotyping. The pooled DNA is assayed on a standard genotyping chip, and allele frequencies are estimated from the raw intensity data for the chip. Pooled genotyping is a cost-effective alternative to individual genotyping in genome-wide association studies. Pooled genotyping can be a useful pre-screening method to select possible markers for a second stage of individual genotyping.

Some previous studies have focused on finding the optimal pooling fraction to generate the optimal pools in order to reduce the cost. They selected some markers under the optimal fraction and generated pools for them to move into linkage/association analysis. Optimal pooling design starts with most individuals ranked by phenotypic trait, and the individuals from the top and the bottom of trait distribution are selected by the optimal fraction rate for the optimal pools. Many previous researchers recommended the 27% optimal pooling fraction rate. Then the optimal pooling design using the optimal fraction rate contained 80% individual genotype information of original data for within-family and between-family association designs (Bader and Sham 2002; Jawaid et al., 2002). However the fraction rate was suggested for common alleles with additive effects with no experimental errors. In case of rare or recessive alleles, the suggested fraction rate was slightly reduced to around 20% (Bader and Sham 2002; Jawaid et al., 2002). However our interest in this study is how to put people into pools, not how many markers are chosen for optimal selection of the next step. Therefore we used a whole list of genes without selecting a few top choices to find out the pooling strategy.

A number of previous studies have looked at the issue of bias in allele frequency estimation from pooled DNA data. Pools are generated by mixing the same amount of DNA from individuals and then amplifying by polymerase chain reaction (PCR) (Hoogendoorn et al., 2000; Le Hellard et al., 2002; Sham et al., 2002). The quantity of each allele in the pool is in different proportion. Therefore it could cause biases during pooling and less accurate pooled allele frequency estimates. There are, however, concerns about both bias and variance of these allele frequency estimates. The bias is due to unequal hybridization of the two alleles and the variance is due to poor measurement of the “equal” quantities of DNA from each person as well as noise in reading the intensity off the chip. A fair amount of literature has addressed the bias issue, and there are a number of papers on correction factors for allele frequency estimates from pooled data. Two main correction factors were proposed for estimating more accurate allele frequency in a pool (Hoogendoorn et al., 2000; Le Hellard et al., 2002; Norton et al., 2002; Simpson et al., 2005). One is called  $k$ -correction factor using the mean of the ratio of the observed two alleles in a heterozygote (Hoogendoorn et al., 2000; Le Hellard et al., 2002; Norton et al., 2002) and the other is modified correction factor using relative allele signal (RAS) values to predict accurate allele frequencies (Craig et al., 2005; Kirov et al., 2006; Simpson et al., 2005). Two RAS values for sense strand (RAS1) and for antisense strand (RAS2) are different measures of the same allele. This is to modify  $k$ -correction factor using RAS values, and then the estimated allele frequencies using the modified correction factor is more accurate allele frequency predictions with less biases (Craig et al., 2005; Norton et al., 2002). The average of two RAS values was also used to derive  $k$ -correction factor, and was used to predict accurate allele frequencies (Simpson et al., 2005). However, as mentioned above, we do not address in the bias

issues, since any bias due to unequal hybridization of allele will apply equally to all DNA samples regardless of phenotype.

The variability issues have also been addressed in a few papers. To recover power that is lost due to increased variability in pooled studies, some authors have considered replication of pools and/or number of samples in a pool. The added variability from DNA pooling could be reduced by using replicate pools either from the same individuals across pools or by dividing the individuals into several pools. And multiple measurement of allele frequency from the same pool also can reduce the measurement error (Le Hellard et al., 2002; Sham et al., 2002; Visscher and Le Hellard 2003; Zou and Zhao 2004). One group suggested using multiple subpools with equal numbers of individuals and triplicate replicates (Pearson et al., 2007). It is generally recommended to use larger pools and to use multiple replicates, but there has not been any systematic study of what kind of replication design is most statistically efficient.

To do any statistical analysis of a pooled study, the design must have multiple pools, since the pool (chip) is the unit of analysis. For any design that meets that criterion, there are a number of choices of analysis methods. Primarily, three kinds of tests are used for a simple comparison of allele frequencies between groups. One is modified two-sample test form with the difference between allele frequencies of two groups as the numerator and the standard deviation of that difference as the denominator (Bader et al., 2001; Kirov et al., 2000; Risch and Teng 1998; Zou and Zhao 2004; Zuo et al., 2006). The difference of RAS values of two groups was also used as the numerator (Pearson et al., 2007). And this type of test statistic is considered to follow an asymptotically normal distribution. Another approach is similar to the previous test, but they considered the squared test statistic. Then they considered the test statistic with a  $\chi^2$ -distribution with 1 df (Bader and Sham 2002; Craig et al., 2005; Sham et al., 2002). Yet another

type of test statistic is a modified  $2 \times 2$  contingency table test (Le Hellard et al., 2002; Visscher and Le Hellard 2003). They considered the test statistic as inflated by the errors for estimating allele frequencies. So the test statistic was modified by the estimates of the sampling variances of the allele frequency under the assumption of no difference between the frequencies. This is called a shrunken version of the classical test statistic, and considered to follow  $\chi^2$ -distribution with 1 df. Bivariate distribution of association test statistic is also suggested to compare the efficiency of pooled genotyping versus individual genotyping (Knight et al., 2009). They set proportions for bad SNPs and good SNPs, and use their joint density. However, the bivariate statistics they suggested are only for SNPs selected from individual genotyping. For testing the difference of allele frequencies, the variations that occurred during the pooling experiment are also considered. Since the errors due to pooling depend on the allele frequency, sample size, pool size, and/or the coefficient of variation of the number of DNA molecules of one locus contributed by each individual,  $\tau$ . All variables except  $\tau$  can be measured easily, but  $\tau$  is believed to have a very small value. In a recent study, an estimate of  $\tau$  is suggested (Jawaid and Sham 2009). These previous test statistics are for simple comparisons between two groups. Since we will consider pooling designs with multiple pools (or chips), we need more complex ANOVA-type models (pooling error and measurement error) and also need to update the test statistics.

In this work we consider both design issues and statistical testing methods. We consider several pooling designs with the same cost (same number of chips) and compare them to figure out which design is more effective. We assume that our pooled study is the first-stage of a genome-wide association study, and that the purpose is to find a list of genes that should be carried to the second stage for individual genotyping. In order to compare designs, we must first

derive correct statistics for hypothesis testing under each design. We also discuss the analysis issue of how our “correct” statistics compare to simpler alternatives.

### **1.4.2 Genotype Calling Methods for Trisomic Samples**

Genotypes of individuals are “called” from raw intensity data by clustering methods. There are a number of different statistical clustering techniques and software packages in use for this, many of them specific to particular genotyping technologies. Most individuals, who are disomic, would have three clusters, but trisomic individuals would have four clusters. For a standard SNP, disomic individuals will have three possible genotypes, which are *AA*, *AB*, or *BB*. However trisomic individuals have four possible genotypes, which are *AAA*, *AAB*, *ABB*, or *BBB*. Standard software packages for genotype calling cannot be used when the individuals being genotyped have a non-standard numbers of chromosomes and thus a non-standard number of clusters. The problem we are interested in is that of genotyping in trisomic individuals – those with three copies of a particular chromosome.

Genotype calling is prerequisite process for any association study. Since it causes errors in the association studies if genotype calls are erroneous, many genotype calling algorithms have been proposed and compared with each other. There are two ways to assign the genotypes (i.e., “call” genotypes): supervised methods and unsupervised methods. Supervised methods are used when a training dataset is available, while unsupervised methods are used when there is a lack of prior knowledge. Some of the popular unsupervised methods are K-means clustering algorithm, hierarchical clustering, and dynamic model-based algorithm (DM) developed for the Affymetrix 100K array. Both K-means and hierarchical clustering algorithms are well-known simple methods, but need additional considerations to handle more complex structures (Kerr et al.,

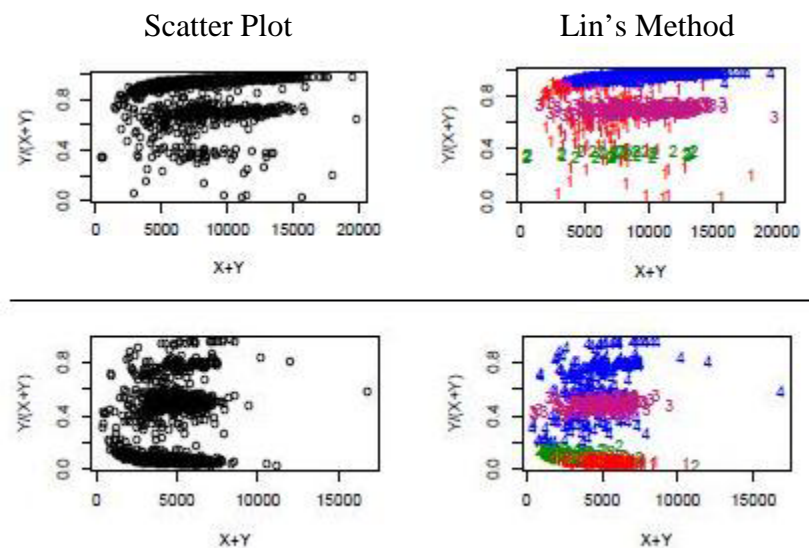
2008). Moreover, the K-means clustering algorithm works well for data that have distinct separate clusters with the same variance in all clusters. However homozygote clusters often have lower variance than heterozygote clusters. Therefore new model-based methods were suggested to enhance the algorithms for the genotype calls regardless of the different variances of clusters. The DM method proposed by Di et al. (2005) is not sensitive to experimental conditions and designs, but is very flexible for balancing call rate and accuracy, so more accurate genotype calling results are presented. Fujisawa et al. (2004) proposed a model-based clustering method using a Gaussian mixture model for data generated with the Invader assay. Since outlier and uncertain genotype detection is important and simple methods like K-means strongly depend on priors, Fujisawa modified the method to be robust to outliers. Recently, Vens et al. (2009) compared three different genotype calling algorithms for the Affymetrix 500k array set - Bayesian robust linear modeling using Mahalanobis distance (BRLMM), Bayesian hierarchical four-class mixture model (Chiamo++), and EM clustering algorithm (JAPL). Then they suggested JAPL to keep as many subjects as possible, and Chiamo++ to use higher number of SNPs for later analysis.

There had not been specific genotype calling methods for trisomic data before Lin et al.(2008). Lin et al. (2008) proposed and compared a modified K-means algorithm, a Gaussian-mixture algorithm, and a novel Beta-mixture algorithm for both disomic and trisomic Illumina data. Both the K-means and the mixture approaches were applied to genotyping individuals and to calling family genotypes as a unit. The family-based likelihoods that Lin et al. (2008) used are similar to the algorithm of Sabatti and Lange (2005), except that Sabatti and Lange used a Bayesian framework. The method of Lin et al. (2008) did not make assumptions about allele frequencies and/or genotype frequencies, and used one-dimensional data after any kind of



transformation. The likelihood-based algorithms by Lin et al. (2008) could apply to data from any platform, and the clusters by Lin's methods are superior to standard clustering methods (eg. K-means) for trisomic individuals. However sometimes the Beta mixture model did not even run, and also it sometimes found the clusters wrong (see Figure 1.2).

Our objective in this study is to improve the genotype calling procedures for various quality levels of raw data, with a particular emphasis on calling genotypes in trisomic samples. Fortunately, there are some aspects of genotype calling to make the process easier (Lin et al., 2008). First, the clusters have quantitative restriction – for instance, three clusters for disomic and four clusters for trisomic. Second, the distribution of the data depends on the genotyping platform the data came from, the data quality, and the transformation used. It could be symmetric and/or skewed distribution according to the platforms even though the data samples are about the same DNA, but this information is known from previous studies.



**Figure 1.2: Wrong clusters by Lin's IBM for trisomic individuals**

## 1.5 PURPOSE OF THIS DISSERTATION

The main purpose of this dissertation is to look at two problems involving raw data from genotyping chips: how to design and analyze pooled DNA association studies, and how to call genotypes for trisomic data.

*Topic 1: Using raw data from pooled studies to test for association.*

In pooled studies, there are issues of both bias and variance that must be dealt with in analyzing the data. Many previous studies have focused on bias issues – how to estimate more accurate allele frequency from genotyping pools. However we contend that the bias issue is not critical, since any bias due to unequal hybridization of alleles will apply equally to all DNA samples regardless of phenotype. Then the bias is almost irrelevant to a hypothesis test of genotype/phenotype association. A number of previous studies have also proposed models for the variability introduced by pooling, but have not necessarily used these models to answer the most pressing questions about test statistics and study design. We use several different models of pooling variability, which are similar to those in previous studies, to consider the following questions:

- 1) What is the most appropriate test statistic for pooled genotype data?
- 2) How does the power compare for the designs we considered?

We considered several pooling designs with the same number of chips, so that all designs would have the same cost. We derived optimal test statistics for each design theoretically, and compared them to find out which design is most statistically efficient. Finally, we compared the performance of various tests statistic options on a real dataset.

We initially considered a simple model without covariates, but most studies of practical interest have at least one covariate like sex, age, and so on. Therefore we considered one additional question, as follows: “How should the study be designed, if we have an important covariate (e.g. sex)?”.

*Topic 2: Genotype calling for trisomic samples.*

Genotype calling methods for trisomic data had not been previously developed until Lin et al. (2008) proposed the procedures for trisomic data. This included genotype calling methods for each person individually, and also for an entire family as a group. They proposed two approaches: modified K-means clustering method and parametric methods. Original K-means clustering method may not perform well if the variances of the clusters are different. Lin et al. (2008) noted that for many assays, the heterozygote group has larger variance than the homozygote groups, and modified K-means method to improve the genotype calls using the pedigree information. Modified K-means updated the centers at the end of iteration step. Similarly Lin et al. (2008) extended parametric methods proposed by Fujisawa et al. (2004) to allow pedigree information to be used to improve the genotype calls using parametric models. Gaussian mixture models have previously been used for this problem, but not Beta as proposed by Lin. They applied these methods to disomic data first, and then extended the method for trisomic data.

In this dissertation, we will review Lin’s genotype calling procedures for trisomic data and will propose a new genotype calling method only for heterozygotes. Lin’s genotype calling method called genotypes by finding four clusters for trisomic individuals. But for Illumina data we have seen that Illumina can find the homozygote clusters very well, so we want to let

Illumina do that. Then our main purpose is to find the middle two clusters well – that is, to find the difference between *AAB* and *ABB* genotypes. We compare Lin’s genotype calling method with the new genotype calling method.

## **1.6 DATASETS**

### **1.6.1 Pancreatic Cancer Pooled Data and Case-Control Study**

This pooling-based case-control study of pancreatic cancer is described in Diergaarde et al. 2009. In this study, the Affymetrix genome-wide human SNP array 6.0 was used in a design that included both subsets of data pooled separately and replicate pools (See Figure 3.3). The case and control groups have the same number of individuals - 62 males and 41 females (total 103 individuals). There are five sub-groups with duplicate pools in each cohort, which consist of 21 males, 21 males, 20 males, 21 females, and 20 females each. Therefore there are five sub-groups with duplicate pools each (total 10 pools) in each cohort. We use this dataset to evaluate the performance of different statistics for testing case-control association with pooled DNA.

### **1.6.2 Down Syndrome Data for Genotype Calling**

Down syndrome is caused by a meiotic nondisjunction event. Nondisjunction is an error that occurs during cell division. This is the failure of homologous chromosomes to separate in meiosis I, or the failure of sister chromatids to separate during meiosis II. If chromosome 21 has the extra chromosome, this is called *Down syndrome* or *Trisomy 21*. We have two real datasets

for Down syndrome, and will use these datasets for testing our trisomic genotype calling methods. All subjects in both datasets were genotyped by Illumina's BeadStudio Genotyping Module. The BeadStudio Genotyping is for analyzing data collected by Illumina's GoldenGate and Infinium genotyping assays.

#### Dataset 1

Dataset 1 consists of 358 SNPs on chromosome 21 genotyped in 262 individuals with Down syndrome. It is a part of a larger dataset (also including genotypes for parents and for some SNPs on other chromosomes) collected as part of a case-control study of atrioventricular septal defects (AVSD) in Down syndrome. Genotyping was done on the Illumina BeadArray platform using the Golden Gate genotyping technology by the Seattle SNPs PGA. The dataset is further described in Locke et al. (submitted).

#### Dataset 2

Dataset 2 consists of 1,536 SNPs on chromosome 21 genotyped in 1,060 individuals with Down syndrome. It is part of a larger dataset that includes the case-control study from Dataset 1 as well as a population-based cohort that is being used to study association between nondisjunction and meiotic recombination. Genotyping was done on the Illumina BeadArray platform using the Golden Gate genotyping technology by the Center for Inherited Disease Research (CIDR).

## **2.0 GENOTYPE CALLING FOR TRISOMIC SAMPLES**

### **2.1 BACKGROUND**

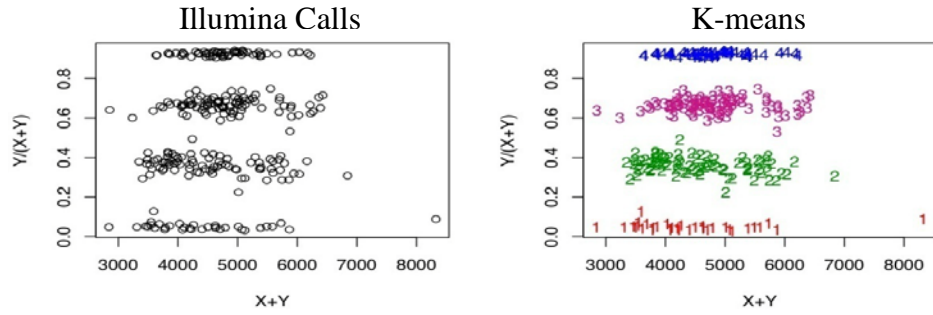
A single nucleotide polymorphism (SNP) is a DNA sequence variation that occurs when a single nucleotide (A, T, C, or G) differs in the genome among a species. SNPs are very popular in contemporary genetic studies, and can be efficiently genotyped in both small and large numbers. Many different technologies can be used for genotyping, but none of them produce a genotype as the primary (raw) form of the data. For all SNP genotyping technologies, genotypes are “called” from raw intensity data using clustering techniques. There are a number of different statistical clustering techniques and software packages in use for this, many of them specific to particular genotyping technologies.

Standard software packages for genotype calling cannot be used, however, when the individuals being genotyped have non-standard numbers of chromosomes. The problem we are interested in is that of genotyping in trisomic individuals – those with three copies of a particular chromosome. Popular clustering methods for genotype calling include K-means, Gaussian model-based, and other variations. While off-the-shelf genotype calling software cannot be applied to trisomic data, the fundamental methods (e.g. K-means) used for disomic calling can be applied to trisomic samples using custom software. We have previously developed such software (Lin et al., 2008) to implement both K-means and model-based clustering using beta

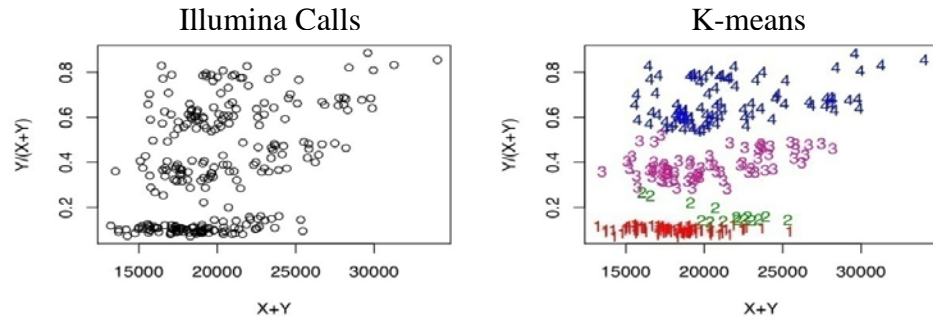
distributions. It is also possible to run standard software on raw data from trisomic samples, and we have found that for several platforms, including Illumina, this produces fairly good results in the sense that homozygotes appear to be called correctly, and the two heterozygote groups, *AAB* and *ABB*, are lumped together and assigned the “*AB*” genotype. In this paper we compare our previous methods for calling trisomic genotypes to a much simpler method that first uses standard software to call heterozygotes vs. homozygotes and then applies a model-based clustering to split the heterozygotes into two clusters.

Some of the general issues that need to be taken into account in genotype calling and in comparing methods are as follows. First, almost all methods work well when clusters are extremely well-separated, as in Figure 1.1 and Figure 2.1-(a). And almost all methods fail to find clusters (as they should) when the assay fails so that there really are no distinct clusters, as in Figure 2.1-(c). So in evaluating genotype calling algorithms, we are really interested in finding the methods that work the best for intermediate data quality, as, for example, in Figure 2.1-(b). “Working the best” can be defined in several ways. The ideal definition would be to match gold-standard genotypes for all individuals, especially for those whose points fall between clusters. But there exist almost no datasets with such gold-standard genotypes available, and certainly none for trisomic data. Another success criterion would be simply to find the clusters correctly for as many SNPs as possible. For example, Figure 2.2 shows a plot where the calling method was clearly unable to correctly find the clusters. Because of the varying SNP quality, all SNPs are not available to use for every calling method. To compare genotype calling methods, Vens et al. (2009) divided all SNPs into several groups according to whether a SNP could be called by what combination of methods, and then used SNPs that passed all methods they wanted to compare.

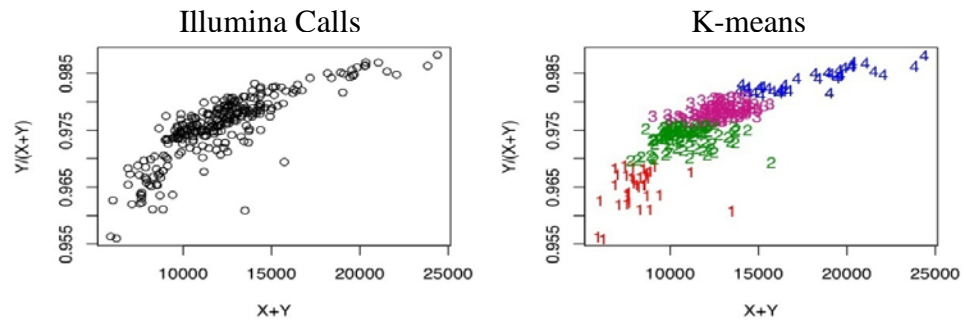
a) SNP with clean distinct clusters



b) SNP with less well-distinct clusters



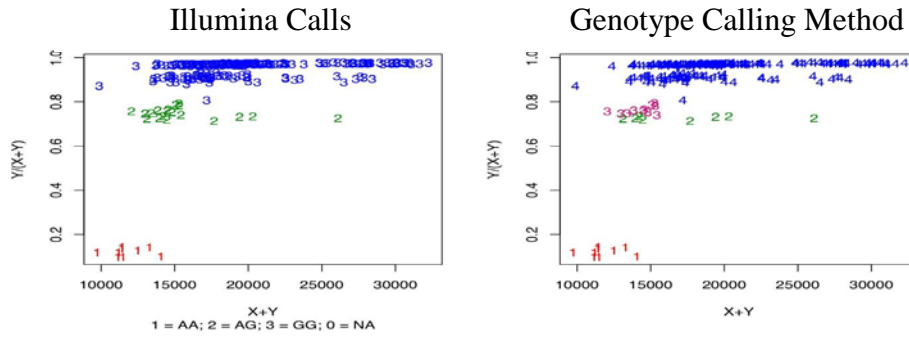
c) SNP with no distinct clusters



**Figure 2.1: Examples of genotype calling results of various SNPs using K-means**



Another important issue in genotype calling is that for large datasets the calling needs to be very automated. It is not possible to manually inspect scatter plots of hundreds of thousands of SNPs, so good methods need to perform reasonably well as consistently as possible without human intervention.

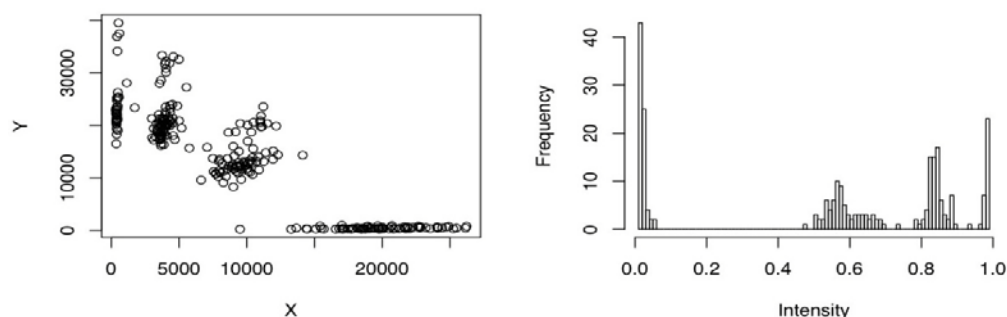


**Figure 2.2: Example of Genotype Clusters Incorrectly Found by Genotype Calling Method**

## 2.2 PREVIOUS GENOTYPE CALLING METHODS

Standard calling methods include both supervised and unsupervised methods, depending on whether training datasets are available. K-means clustering is one of the popular unsupervised methods. This algorithm performs well for high-quality data with well-separated clusters, but it can perform poorly when the variances of the clusters are not the same. Homozygote clusters often have lower variance than heterozygote clusters, and in that situation model-based clustering methods can do better, since they can estimate different variances for each cluster. For example, a Gaussian-mixture model for Invader assay data was proposed by Fujisawa et al.(2004). In addition, the number of clusters must be known to use K-means, but a mixture

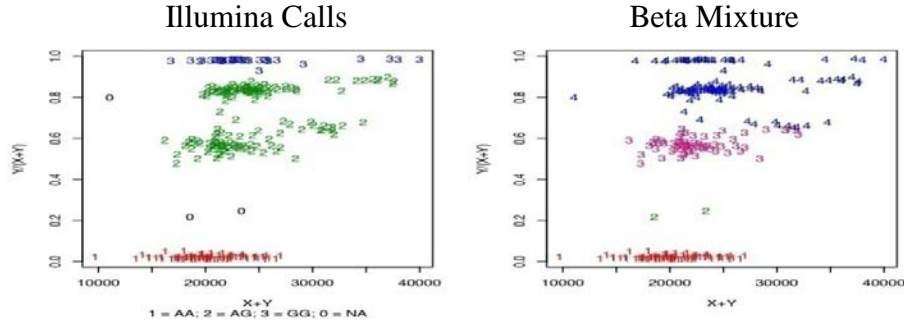
model can use a penalized likelihood method for finding the correct number of clusters (Lin et al., 2008). Lin et al. (2008) was the first to propose a beta-mixture model, which is able not only to estimate separate variances for each cluster, but also to model skewness of the homozygote clusters (see Figure 2.3). Modeling the skewness reduces the number of miscalls as compared to Gaussian models. Lin et al. (2008) developed versions of the beta mixture model for both disomic and trisomic data. If the individuals being genotyped include families, then pedigree information can also be incorporated into genotype calling (Lin et al., 2008; Sabatti and Lange 2005), but we do not consider family data in the current paper.



**Figure 2.3: Skewness of Homozygote Clusters**

The beta mixture model works well, especially for trisomic data (Lin et al., 2008), but the model is harder to fit than a Gaussian model, and in practice it sometimes fails to run or fails to find the right clusters. By contrast, in our experience standard disomic genotype calling methods (e.g. as implemented in Illumina’s BeadStudio software) are typically fairly robust in their ability to find clusters and identify both heterozygote clusters as “AB”. Figure 2.4 shows one example in which the beta mixture model fails to find the correct clusters, while the BeadStudio calls correctly distinguish heterozygotes from homozygotes. This observation motivates the current

work. Is it possible to improve trisomic genotype calling by starting with the disomic calls and then “splitting” the heterozygote cluster in two?



**Figure 2.4: Examples of genotype calling results for trisomic individuals with misclassified clusters by Lin’s IBM**

## 2.3 OUR PROPOSED METHOD FOR TRISOMIC GENOTYPE CALLING

Our new method starts with data from any standard disomic genotype calling algorithm, and then uses a Gaussian mixture model to split apart the heterozygotes into  $AAB$  and  $ABB$  clusters. The Gaussian model suffices here (as opposed to the beta), because the two heterozygote clusters are expected to have relatively similar and symmetrical distributions. We did however, experiment with using K-means to split the heterozygotes, and found that it did not perform as well for SNPs with less distinct genotype groups. We refer to our method as the Heterozygote Independent Normal Mixture (HINM) procedure. The HINM method can be applied only to trisomic individuals, not to disomic individuals. HINM makes the explicit assumption that the disomic calls for the homozygote individuals are correct. That is, we assume that individuals with  $AA$  disomic calls are actually  $AAA$ , those with  $BB$  calls are  $BBB$ , and those with  $AB$  calls are to be

classified into two genotype groups (*AAB* or *ABB*). Thus it should be noted that the HINM method automatically fails if the original disomic calls are poor.

The concept of HINM is similar to that proposed by Lin et al. (2008), but we are able to use a simpler model since only heterozygotes need to be considered. Let  $y$  be the observed value for an individual, which follows parametric model.

$$y/g = \lambda \sim f(\xi_\lambda),$$

where  $\lambda \in A = \{AAB, ABB\}$  and  $f(\xi_\lambda)$  denotes any parametric model with  $\xi_\lambda$ , being a parameter vector for genotype  $\lambda$ . For heterozygote individuals, in most SNPs, these values are usually fairly symmetrically distributed. Therefore we take  $f(\xi_\lambda)$  as a Gaussian-mixture model with parameter,  $\xi_\lambda = (\mu_\lambda, \sigma_\lambda^2)$ . Let  $y_i$  and  $g_i$  be the observed value for  $i$ -th heterozygote individual and the corresponding genotype, respectively. Then the likelihood for the  $i$ -th heterozygote individual would be

$$L_i(y_i, g_i, \xi_i) = \Pr(g_i) \Pr(y_i|g_i) = \prod_{\lambda \in A} \Pr(g_i) f(y_i, \xi_i)^{1_{\{g_i=\lambda\}}},$$

where  $\xi_i$  is a parameter of the  $i$ -th individual. The likelihood for  $n$  heterozygote individuals would be

$$L(y, g, \xi) = \prod_{i=1}^n L_i(y_i, g_i, \xi_i).$$

We can compute the probabilities of genotypes using Bayes' rule when parameters are known. The posterior probability of the genotype for heterozygote individuals given their observed values is

$$P(g|y) = E/F,$$

where  $E = \Pr(g_i) \Pr(y_i | \zeta_{\lambda=g_i})$  and  $F = \sum_{j=1}^n \Pr(g_j) \Pr(y_j | \zeta_{\lambda=g_j})$ . And we can get the estimated parameters using an expectation maximization (EM) algorithm. The EM update expressions for HINM are

$$p_{\lambda}^{(t+1)} = \frac{E(S_{1,\lambda} | y, \theta^{(t)})}{2n}$$

$$\mu_{\lambda}^{(t+1)} = \frac{E(S_{2,\lambda} | y, \theta^{(t)})}{E(S_{1,\lambda} | y, \theta^{(t)})}$$

$$\sigma_{\lambda}^{2(t+1)} = \frac{E(S_{3,\lambda} | y, \theta^{(t)})}{E(S_{1,\lambda} | y, \theta^{(t)})} - (\mu_{\lambda}^{(t+1)})^2,$$

where  $S_{1,\lambda} = \sum_{i=1}^n 1\{g_i = \lambda\}$ ,  $S_{2,\lambda} = \sum_{i=1}^n 1\{g_i = \lambda\} y_i$ , and  $S_{3,\lambda} = \sum_{i=1}^n 1\{g_i = \lambda\} y_i^2$  with the parameter  $\theta = (p_{\lambda}, \mu_{\lambda}, \sigma_{\lambda}^2)^T$ . To get initial values for the means and variances we first cluster using K-means and calculate the means and variances of those clusters. We set the probabilities of genotype,  $p_{\lambda}$ , to be 0.5 as initial values.

## 2.4 EXAMPLE DATASETS

We test our methods on two example datasets of individuals with trisomy 21 (Down syndrome). Dataset 1 consists of 358 SNPs genotyped in 262 individuals, and Dataset 2 consists of 1,536 SNPs genotyped in 1,060 individuals. More detailed data descriptions are in Chapter 1.6.2. Both datasets were genotyped using the Illumina Golden Gate technology, but genotyping was performed in different genotyping centers. Genotypes were called in both datasets using Illumina's BeadStudio software, but using site-specific protocols for settings, hand-adjustments, etc.

## 2.5 COMPARISON OF METHODS

We applied Lin et al.'s beta mixture methods and our HINM method to each dataset using the following procedure.

- A. *Score the data by quality*: Since we are interested in how the methods perform for SNPs with good clusters, poor clusters, and medium-quality clusters, we started by hand-inspecting and scoring each SNP by quality. We made scatter plots of transformed intensity for every SNPs individually. Based on the scatter plots, we assigned each SNP a quality score from 0 to 3 – 3 for nice clusters; 2 for nice clusters with some low-intensity points; 1 for muddy clusters; and 0 for no clusters/monomorphic cluster. These scores were used only for the purposes of reporting results; they were not used in any way during the genotype calling process.

- B. *Find a set of SNPs for which both methods run:* We initially applied the two genotype calling methods to every SNP in each dataset. Overall the failure rate for Lin's IBM is over two-fold higher than for HINM (detailed in Chapter 2.6.2) for both datasets. For most SNPs with quality 0 in dataset 1, the success rate of both methods is extremely low. And for SNPs with quality 1 in dataset 1, the success rate of both methods is low as well. However most SNPs in dataset 2 work with both methods regardless of the SNP quality. We do not know the exact reason, but we believe that a major reason for this difference is sample size; the number of individuals in dataset 2 is about three times the number in dataset 1.
- C. *Choose a set of SNPs for comparison:* For the comparison, we randomly selected 200 SNPs in each dataset from the set of SNPs for which both genotype calling algorithms ran.
- D. *Compare results:*
- (a) Count mismatch calls in each SNP.

We use mismatch calls between Lin's IBM and HINM methods among the selected 200 SNPs. However missing Illumina calls are excluded for the comparison. Among 200 SNPs, 112 SNPs in dataset 1 and 1 SNP in dataset 2 have exactly the same call results for every SNP by Lin's IBM and HINM. However 88 SNPs in dataset 1 and 199 SNPs in dataset 2 have at least one mismatch. These were used for the comparison.

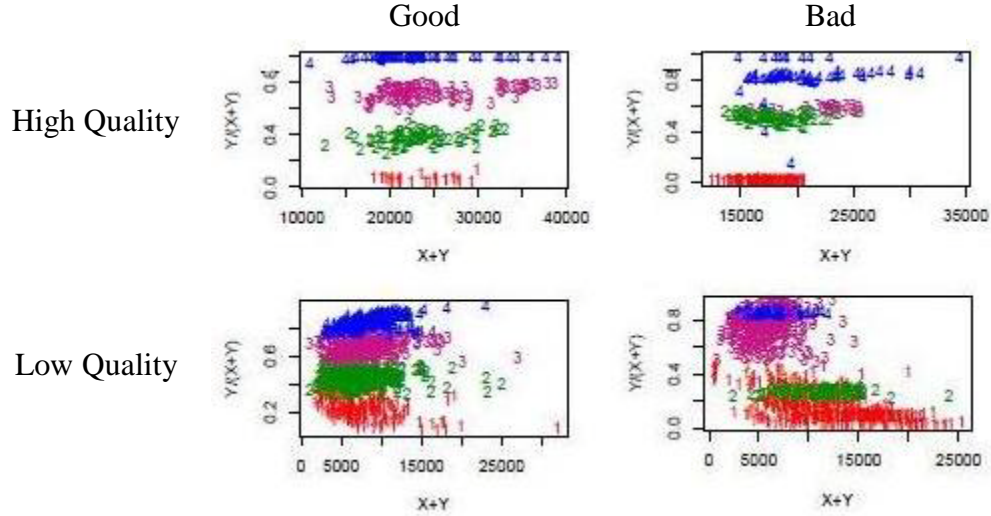
- (b) Define the groups based on mismatch rate (*MMR*).

Since mismatch rate is from 0.03% to around 90% in both datasets, we chose 4 datum points (10%, 30%, 50%, and 70%) and divided into 5 groups. Among 5 MMR groups, the call results by Lin's IBM and HINM are similar with each other if MMR is 1, and it is hard to say which method worked better. And it is not useful to compare the methods. Therefore to compare the call results, we only consider SNPs if MMR is 2 to 5.

(c) Compare the methods.

To compare the calls, we classified calls into 4 categories by visual inspection – HINM better, IBM better, Good for both, Bad for both. Figure 2.5 shows examples of Good and Bad calls for high and low quality SNPs. These categories are defined by comparing scatter plot with call plot. If there are 4 visible clusters in scatter plot, it is easy to classify good or bad calls based on the call results. High quality SNPs usually have 4 clusters by eye detection in scatter plot, and both methods worked pretty well to identify the clusters. However for low quality SNPs with muddy clusters, there is no standard for classifying good or bad call results even though the method worked. Therefore we classified low quality SNPs into “good” call results, if the call clusters are well-separated (but usually they are close each other) and in order. On the other hand, if the genotype groups are intermixed or out of order, we classified the SNP into “bad” call results. If calls by both methods are good, the SNP is in “Good for both” but it is hard to say which method is better. On the contrary, if the calls by both methods are bad, the SNP is in “Bad for both”. After categorizing the call results, we compare the number of SNPs by MMR, SNP quality, and categorized call results.





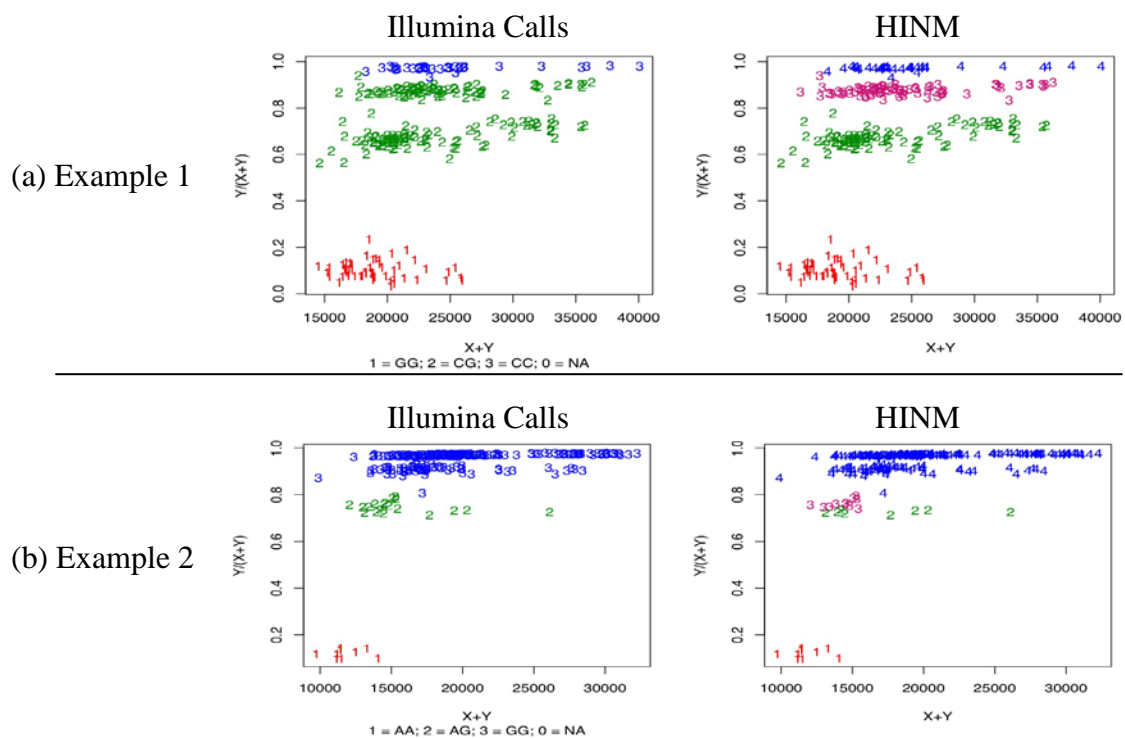
**Figure 2.5: Example of "Good" and "Bad" calls of high and low quality SNPs**

## 2.6 RESULTS

### 2.6.1 Genotype Calling by HINM

HINM has focused on classifying the heterozygote Illumina group into two groups. And the two homozygote Illumina groups directly interpreted as homozygote groups by HINM. Figure 2.6 shows how HINM worked for two example SNPs. In the Illumina call plot, “1” and “3” denote homozygote samples, and “2” denotes heterozygote samples. On the other hand, in the HINM genotype call plot, “1” and “4” represent homozygote groups, and “2” and “3” represent heterozygote groups. The SNP in Figure 2.6-(a) has clear distinct four groups on the Illumina genotype plot. For this SNP, the HINM method called the genotypes well. The SNP in Figure 2.6-(b) also has four distinct groups on the Illumina genotype plot, however three clusters (two

upper clusters and lower cluster) are for individuals with homozygote Illumina genotypes. Therefore the HINM method follows to call the three groups for homozygote individuals as homozygote individuals, and only one group left with heterozygote Illumina genotypes is classified into two heterozygote individual groups. This is classified as a bad calling result.



**Figure 2.6: Examples of genotype calls by HINM**

## 2.6.2 Comparison of Results of HINM with Lin's IBM

### 2.6.2.1 Apply both HINM and Lin's IBM methods

Both methods were applied to Dataset 1 (262 individuals with 358 SNPs) and Dataset 2 (1,060 individuals with 1,538 SNPs). Among 358 SNPs in Dataset 1, Lin's IBM and HINM worked for

236 SNPs (66%) and 295 SNPs (82%), respectively. Both methods worked simultaneously on 231 SNPs. Only HINM is applicable to call genotypes for 64 SNPs, while 5 SNPs worked with Lin's IBM only. On the other hand, among 1,536 SNPs in Dataset 2, Lin's IBM and HINM called the genotypes for 1,457 SNPs (95%) and 1,525 SNPs (99%), respectively. 1,454 SNPs were called by both methods. While 71 SNPs worked by HINM only, only 3 SNPs worked by Lin's IBM only.

**Table 2.1: Number (Percentage) of SNPs of each quality in each dataset**

SNP Quality	0	1	2	3	Total
Dataset 1	58 (15.6%)	88 (20.9%)	92 (30.7%)	120 (32.6%)	358 (100%)
Dataset 2	121 (7.8%)	162 (10.5%)	48 (3.1%)	1,205 (78.5%)	1,536 (100%)

**Table 2.2: Percentage of SNPs that could be called with each algorithm**

Method	SNP Quality Dataset	0	1	2	3	Total
IBM	Dataset 1	0.8%	7.5%	26.8%	30.7%	65.9%
	Dataset 2	7.7%	10.1%	2.9%	74.1%	94.8%
HINM	Dataset 1	2.5%	17.0%	30.2%	32.7%	82.4%
	Dataset 2	7.9%	10.5%	3.1%	77.8%	99.3%
Both	Dataset 1	0.8%	6.4%	26.5%	30.7%	64.5%
	Dataset 2	7.7%	10.1%	2.9%	74.0%	94.7%

Table 2.1 shows the number (percentage) of SNPs of each quality in each dataset, and Table 2.2 shows the percentage of SNPs that were called by each method. We found that the data quality is related to the sample size and also affects the performance of the genotype calling method. Based on Table 2.2, HINM looks less sensitive to data quality for genotype calling than Lin's IBM method. To compare the methods, as mentioned above, we consider 200 randomly selected SNPs from each dataset, for which both methods (Lin's IBM and HINM) work simultaneously. Table 2.3 shows the number of SNPs among the selected 200 SNPs by quality in each dataset.

**Table 2.3: Number of randomly selected 200 SNPs worked for both methods**

SNP Quality	0	1	2	3	Total
Dataset 1	5	35	62	98	200
Dataset 2	16	33	65	86	200

### 2.6.2.2 Count Mismatches

Among 200 SNPs, we selected SNPs with at least one mismatch, and divided into 5 MMR groups as described in Chapter 2.5. Table 2.4 shows the numbers of SNPs by MMR and SNP quality. Most of the SNPs with mismatches are in the 1<sup>st</sup> MMR group whose mismatch rates are less than 10% (61.4% of 88 SNPs in dataset 1; 75.9% of 199 SNPs in dataset 2). And many high quality SNPs in both datasets also belong to the 1<sup>st</sup> MMR group. As mentioned in Chapter 2.5, only SNPs in from the 2<sup>nd</sup> to 5<sup>th</sup> MMR groups are considered for comparison. Since most of the SNPs in the 1<sup>st</sup> MMR group have similar call results, they are not useful for comparison of the methods.

**Table 2.4: Number of mismatch SNPs by mismatch rate (MMR) and SNP quality**

Dataset 1						Dataset 2					
Quality MMR	0	1	2	3	Total	Quality MMR	0	1	2	3	Total
1	3	8	23	20	54	1	4	11	54	82	151
2	0	6	0	1	7	2	5	14	7	2	28
3	2	8	1	1	12	3	3	3	2	1	9
4	0	6	4	0	10	4	2	5	2	0	9
5	0	3	2	0	5	5	2	0	0	0	2
Total	5	31	30	22	88	Total	16	33	65	85	199

For SNPs from the 2<sup>nd</sup> to 5<sup>th</sup> MMR groups, we categorized into 4 categories by call results and counted them by SNP quality and/or mismatch rates (MMR). Table 2.5 shows the number of SNPs of SNP quality and/or MMR by categorized call results. For both datasets 1 and 2, the HINM method worked better than Lin's IBM method regardless of SNP quality and mismatch rates (67.6% of 34 SNPs in dataset 1; 52.1% of 48 SNPs in dataset 2). For dataset 1, Lin's IBM method also worked well especially for low quality SNPs (quality 0~1) and for SNPs with lower mismatch rate (MMR <50%), but didn't have better performance than the HINM method. When both methods failed to call clearly, the SNPs generally have lower quality (0 or 1). For low quality SNPs with muddy clusters (quality 0 or 1) in scatter plot, HINM method worked better. And for high quality SNPs (quality 2 or 3), HINM method also worked better.

**Table 2.5: Number of SNPs of quality and/or mismatch rate (MMR) by categorized call results for both Dataset 1 and Dataset 2**

**Dataset 1:**

Quality Results \	0	1	2	3	Total
HINM better	1	13	7	2	23
IBM better	1	7	0	0	8
Good for both	0	2	0	0	2
Bad for both	0	1	0	0	1
Total	2	23	7	2	34

MMR Results \	2	3	4	5	Total
HINM better	3	8	8	4	23
IBM better	3	3	1	1	8
Good for both	0	1	1	0	2
Bad for both	1	0	0	0	1
Total	7	12	10	5	34

**Dataset 2:**

Quality Results \	0	1	2	3	Total
HINM better	4	9	10	2	25
IBM better	0	0	0	0	0
Good for both	6	7	1	1	15
Bad for both	2	6	0	0	8
Total	12	22	11	3	48

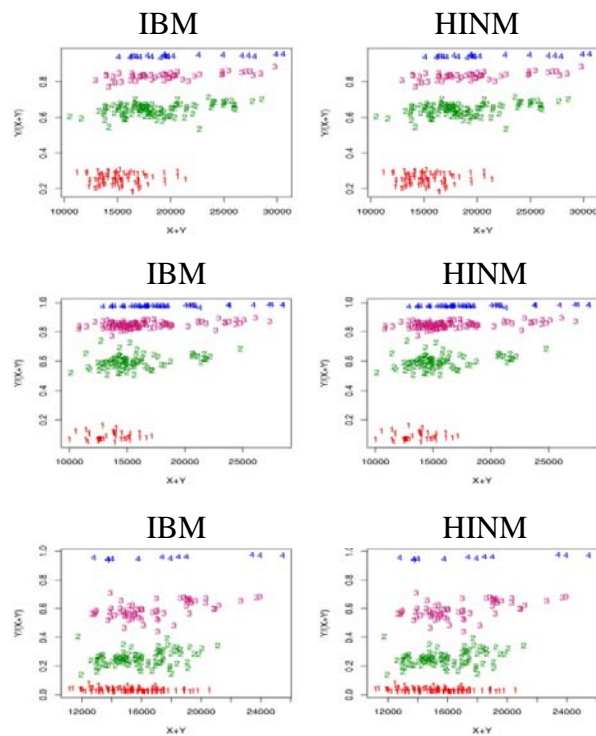
MMR Results \	2	3	4	5	Total
HINM better	12	5	7	1	25
IBM better	0	0	0	0	0
Good for both	10	2	2	1	15
Bad for both	6	2	0	0	8
Total	28	9	9	2	48

### 2.6.2.3 Examples for Visual Detection

Figure 2.7 shows examples of good genotype call results by both Lin's IBM and HINM methods. The first two columns (left-hand-side) are results of SNPs from Dataset 1, and the last two columns (right-hand-side) are results of SNPs from Dataset 2. The first and third columns are by Lin's IBM method, and the second and forth columns are by HINM method. In all the plots, "1" and "4" represent homozygote groups, and "2" and "3" represent heterozygote groups. All SNPs in Figure 2.7 have quite high-quality genotype clusters, and all four transformed clusters in each SNP are almost parallel. Then both Lin's IBM and HINM call the samples well, and the genotype call results are almost the same, except for just a few samples.

Ideally we expect parallel genotype clusters in the scatter plots when the raw intensities are transformed. However sometimes there is a curvature of low-intensities (sum of x- and y-intensities). Among the selected 200 SNPs, 62 SNPs in Dataset 1 (31%) and 65 SNPs in Dataset 2 (32.5%) have low-intensity curvature. Usually homozygote groups have more severe curvature patterns than heterozygote groups. One homozygote group with high-intensities ( $y\text{-intensity} / \text{sum of x- and y-intensities}$ ) usually goes up to close to 1 when sum of x- and y-intensities is getting larger. In contrast with this, the other homozygote group with low-intensities ( $y\text{-intensity} / \text{sum of x- y-intensities}$ ) usually goes down to close to 0. Both methods worked pretty well for these SNPs, however usually call results by both Lin's IBM and HINM methods are different, especially for homozygote genotype groups with the low-intensity curvature points. Figure 2.8 shows genotype calls for pretty good quality SNPs with low-intensity curvature.

a) Dataset 1



b) Dataset 2

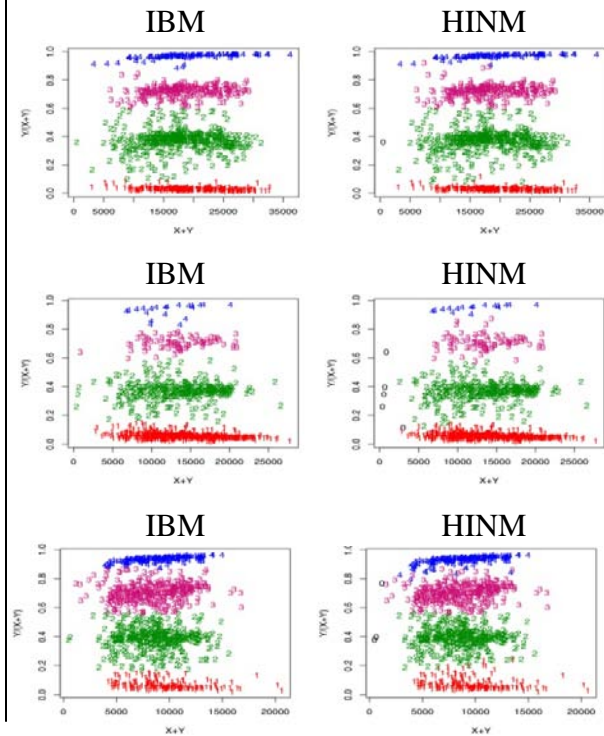
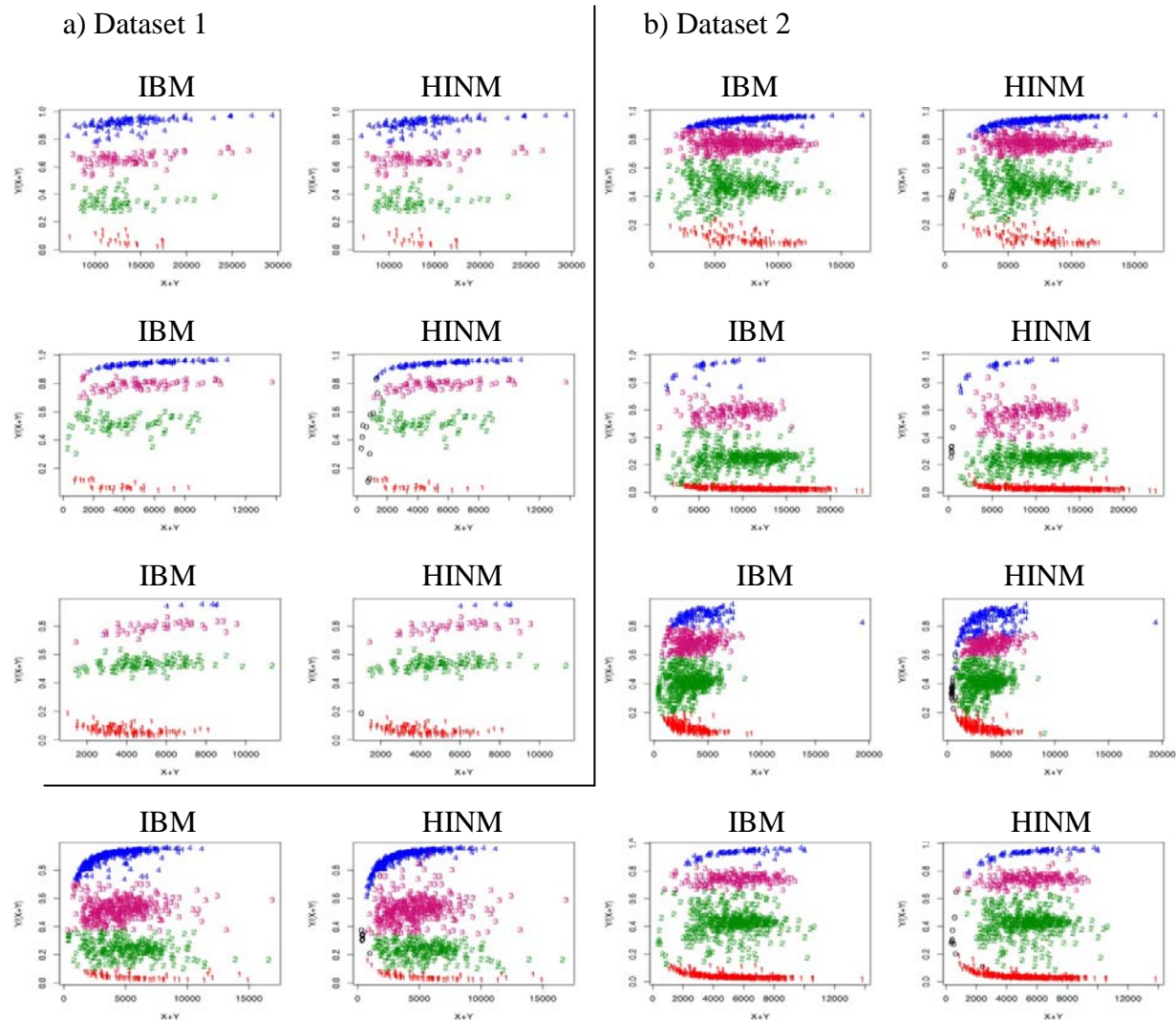


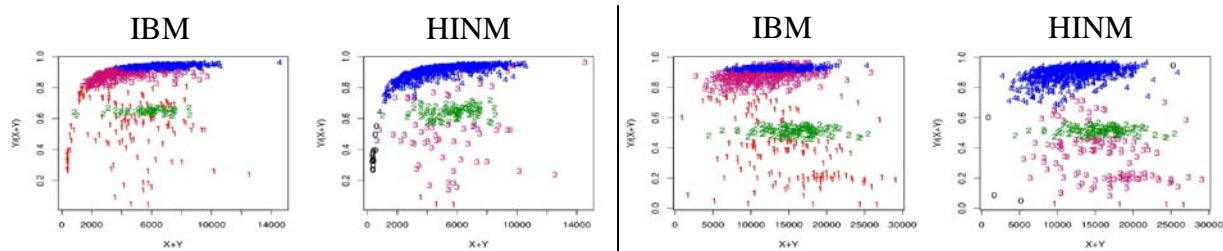
Figure 2.7: Good call results of both Lin's IBM and HINM





**Figure 2.8: Genotype calls for SNP with low-intensity curvature by Lin's IBM and HINM**

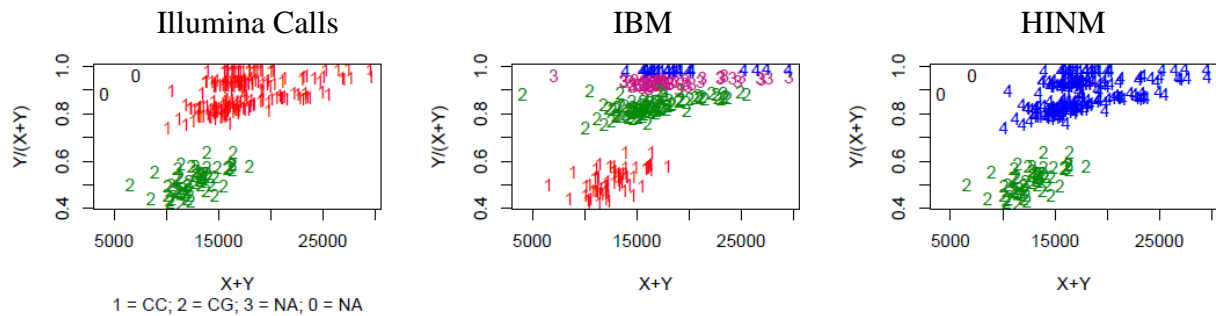
Both Lin's IBM and HINM methods cannot always call the genotypes well. Though trisomic individuals should have four genotype clusters theoretically, some SNPs do not. For these SNPs, both Lin's IBM and HINM methods failed to call genotypes correctly. Figure 2.9 shows some other types of failure examples of genotype calling from Dataset 2. Figure 2.9 shows that SNPs with low-intensity curvature and/or low-quality SNPs are not always assigned the genotypes reasonably. Sometimes all genotype calling methods failed. Even if any of methods succeed to call the genotypes, the genotype call results could be clearly incorrect.



**Figure 2.9: Bad genotype calls for both Lin's IBM and HINM**

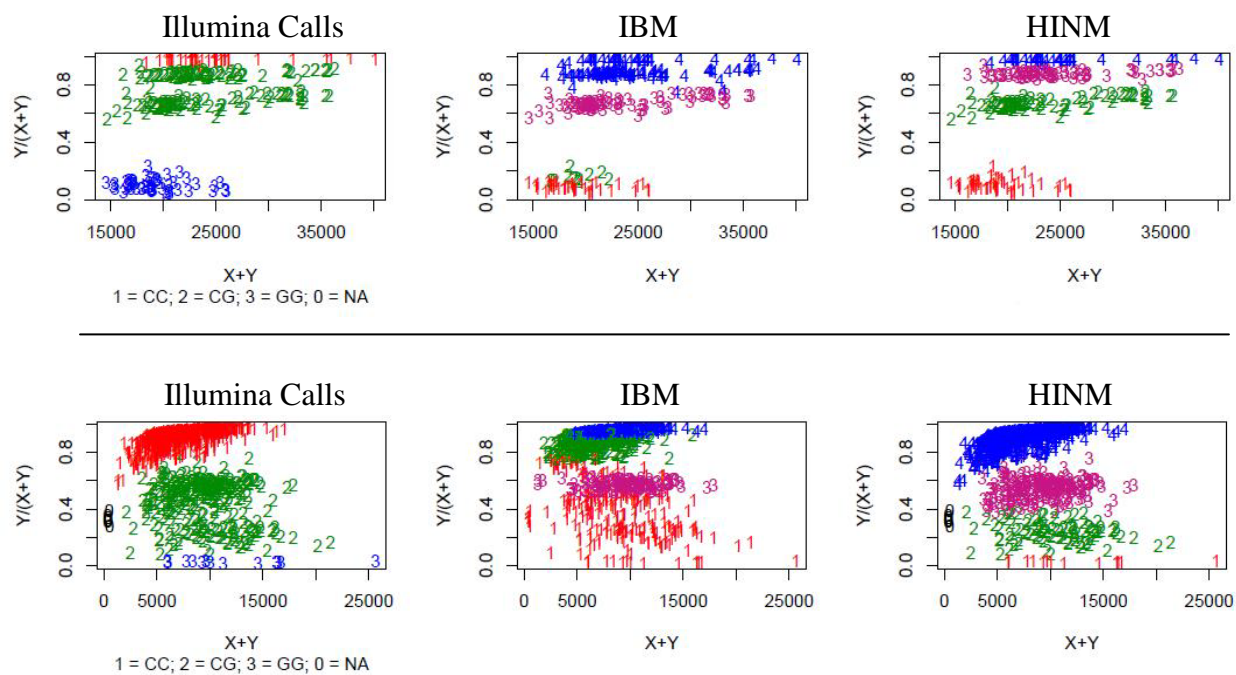
For some SNPs, Lin's IBM method called the genotypes more correctly. Among SNPs with at least one mismatches in Table 2.5, 8 SNPs in Dataset 1 (23.5% of 34 SNPs) and 0 SNPs in Dataset 2 (0%) have better classified genotype clusters by Lin's IBM. Figure 2.10 shows examples of better genotype calls by Lin's IBM than HINM in datasets 1 and 2. In general, these are the cases in which Illumina genotypes are not correct at the first. In this figure, there are four clusters in the plot and Lin's IBM method called them correctly. However, in the Illumina call plot, we can see that one homozygote group has 3 clusters. The HINM method called these

genotypes heterozygotes under the assumption that Illumina already called them correctly. Therefore the HINM method called one homozygote Illumina with 3 clusters as one group.



**Figure 2.10: Better genotype calls by Lin's IBM than HINM**

There are many SNPs for which HINM method called the genotypes better than Lin's IBM method. Among SNPs with at least one mismatches in Table 2.5, 23 SNPs in Dataset 1 (67.6% of 34 SNPs) and 25 SNPs in Dataset 2 (52.1% of 48 SNPs) have well-classified clusters by HINM method based on the plots. Figure 2.11 shows some example SNPs with better genotype call results by the HINM method. The first SNP looks to have four clusters. However Lin's IBM miscalled two clusters (one homozygote Illumina genotype group and one heterozygote Illumina genotype group) as one group when two clusters are close to each other. The second SNP looks to have four clusters with some questionable points. Lin's IBM method failed to call the unclear points correctly, while HINM called the genotypes more clearly.



**Figure 2.11: Better genotype calls by HINM than Lin's IBM**

## 2.7 CONCLUSION

Many algorithms for genotype calling have been developed and used, especially for disomic individuals. However any calling method needs to be modified when a chromosomal abnormality has occurred. We focused on trisomic individuals with Down syndrome (trisomy 21). In a previous study, Lin et al. (2008) suggested a parametric mixture algorithm under the normal or beta distribution to call the genotypes. However in many case of SNPs with unclear clusters and low-intensity curvature, Lin's method failed to run. Therefore we consider a parametric mixture algorithm like Lin's in a little different way under an assumption – that homozygote Illumina calls are perfect. And we applied a parametric mixture model to

heterozygote Illumina calls, and this method is called HINM. After applying both methods, we compared them using mismatch rate (MMR), SNP quality and genotype call results. Genotype call results are classified by our standard of good or bad result as described in Chapter 2.5. The easiest way to check the genotype call result is to compare with scatter plots visually. This standard is not the only possibility – other definitions are possible. Performance for bad SNPs is much less important than the performance for good SNPs. We only use SNPs with mismatch rate over 30% for comparison, since SNPs with low mismatch rates do not contribute much information. Overall the HINM method worked better than Lin's IBM method regardless of SNP quality (0~3) and mismatch rates (MMR). A few cases of failure to run both methods occurred for low quality SNPs. In the case of SNPs with low-intensity curvature, they are clustered pretty well usually by HINM method. Therefore any genotype calling method is useful to call the genotypes for high-quality SNP data, but for low-quality SNP data the HINM method would be better to call the genotypes than Lin's method.

### **3.0 MODELS, TEST STATISTICS, AND DESIGNS FOR GENETIC ASSOCIATION STUDIES WITH POOLED GENOTYPING**

#### **3.1 INTRODUCTION**

Genome-wide association (GWA) studies are now a standard tool for genetic epidemiology, despite their high cost. Almost all GWA studies are done by individual genotyping of all samples in both original and replication datasets, although some recent studies have increased efficiency by using pooled genotyping at early stages (Diergaarde et al., (in press); Nakabayashi et al., 2009; Pearson et al., 2007; Sham et al., 2002; Tabeta et al., 2009; Zuo et al., 2006). There have been several side-by-side comparisons of individual and pooled genotyping that have shown pooled genotyping to be efficient and effective (Bader et al., 2001; Knight et al., 2009; Zou and Zhao 2004), but a critical barrier to more widespread use of pooled genotyping has been the fact that pooling introduces both bias and variance into allele frequency estimates, and this has been perceived as an insurmountable hurdle to performing credible and replicable studies. In fact, however, classical statistical approaches are available to address the bias and variance issues, and the more general credibility issue is typically addressed by replicating results in additional populations (and with individual genotyping) anyway. The purpose of this paper is to suggest pooling designs and statistical tests that are appropriate for the initial screening stage of GWA studies using pooled genotyping.

In a pooled-genotyping association study, DNA pools are constructed by mixing equal amounts of DNA from multiple individuals and then assaying the pool on a single genotyping chip. In older studies, it was not uncommon to pool all cases on one chip and all controls on another, but it is clear that this design does not allow for statistical comparison, since the effective sample size is one. More contemporary pooled studies have divided cases and controls into small groups and used one chip for each group, or have done technical replicates (multiple chips per pool) or both (Bader and Sham 2002; Diergaarde et al., (in press); Jawaid et al., 2002; Sham et al., 2002). It is clear that all of these types of replication can help reduce the extra variability introduced by pooling (e.g. Sham et al., 2002), but there has been little literature that we are aware of comparing the efficiency of different pooling designs. In this paper we consider several standard pooling designs and explicitly compare their statistical efficiency under several different pooling models.

Once genotyping is completed, the data for each SNP on each chip consist of intensity values for each allele, which must be combined in some way to produce an estimate of the allele frequency for the pool. This process is somewhat platform-dependent, but in general involves calculating a ratio of the intensity of one allele to the total intensity for the SNP. On the Illumina platform, the usual terminology is "B allele frequency," and on the Affymetrix platform one often sees "relative allele signal" (RAS), but in both cases these may be calculated in any of a number of different ways. Because of unequal hybridization efficiencies of the two alleles, the intensity ratio derived from the raw data is almost always a biased estimate of the allele frequency. The bias is specific to the SNP and the platform. A number of authors have suggested bias corrections and even set up libraries of correction factors for common genotyping platforms (Craig et al., 2005; Hoogendoorn et al., 2000; Le Hellard et al., 2002; Norton et al., 2002;

Simpson et al., 2005). In this paper, however, we take the position that bias correction is unnecessary if the goal is hypothesis testing of cases vs. controls, since any unequal hybridization efficiency will apply equally to all samples regardless of phenotype. We focus instead on issues of variability introduced by pooling.

Statistical testing of cases vs. controls at each SNP can be done in several different ways. The intensity ratios can be treated as generic continuous outcomes, and the case chips compared to the control chips using a z-test or t-test or something similar (Bader et al., 2001; Diergaarde et al., (in press); Kirov et al., 2000; Pearson et al., 2007; Risch and Teng 1998; Zou and Zhao 2004; Zuo et al., 2006). These tests can use a standard t-test denominator, or they can use a denominator (standard error estimate) that is based on a model for the errors introduced by pooling. Instead, if intensity ratios are considered to be allele frequencies, they can be treated as if they were derived from individual genotyping and a chi-squared test can be performed (Bader and Sham 2002; Craig et al., 2005; Le Hellard et al., 2002; Nakabayashi et al., 2009; Sham et al., 2002; Tabeta et al., 2009; Visscher and Le Hellard 2003). It is clear that both the chi-squared test and the standard t-test would be anti-conservative, since they do not fully account for the pooling variability, but it is not clear exactly how they compare to each other and to the t-test variants that are based on a pooling model. We address that question in this paper.

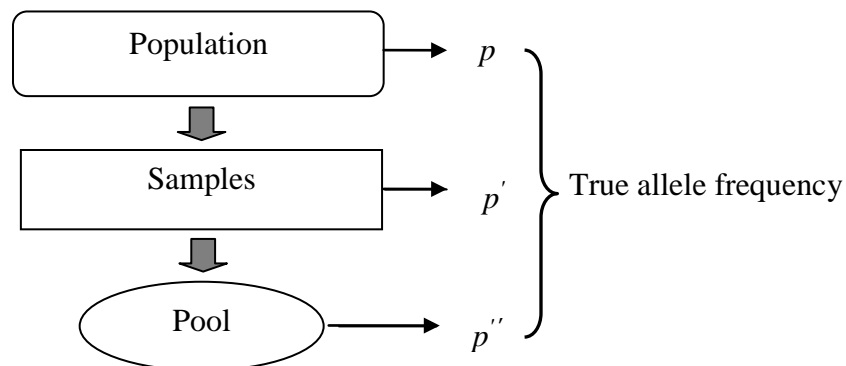
Previous literature on statistical issues in pooled genotyping has developed good models for the data, but it has not really applied those models to answer the most pressing questions about test statistics and study design. In this paper we look at pooling models that are similar to those in the previous literature, and use them to answer the following questions: 1) What is the most appropriate test statistic for a case-control comparison using pooled genotype data? 2) How does



the efficiency of standard designs compare? In the discussion we also consider the issue of designs that incorporate covariates.

### 3.2 MODEL OF POOLING VARIABILITY

Let the true allele frequency of a particular allele in the population be  $p$ , the true allele frequency in a sample of  $N$  individuals be  $p'$ , and the true allele frequency in a DNA pool created by sampling DNA from those individuals be  $p''$  (See Figure 3.1).



**Figure 3.1: True allele frequencies of population, samples, and pool**

Between  $p$  and  $p'$  we introduce variability via standard binomial sampling. That is, the distribution of  $p'|p$  is binomial  $(2N, p)$ . Between  $p'$  and  $p''$  we introduce error via a more complex binomial sampling process, as follows (Jawaid et al. 2002). For each individual in the mixture, we sample a large number of DNA molecules. This number is approximately the same for each person, but is most appropriately modeled as a random variable. Let  $X_i$  be the total molecules

(alleles) sampled from individual  $i$ . We model  $X_i$  as normally distributed with mean  $\mu$  and variance  $\tau^2 \mu^2$ , where  $\tau$  is the coefficient of variation of the number of DNA molecules sampled. It is well established in the lab that the variance of  $X_i$  does depend on the mean, but it is also believed that if the DNA quantification is done properly that  $\tau$  should be quite small - typically less than 0.1 (Jawaid and Sham 2009). Let  $Y_i$  be the total number of A alleles sampled from individual  $i$ . Then  $Y_i$  has density depending on  $X_i$  and on the actual genotype of person  $i$ , as follows.

$$Y_i = \begin{cases} X_i & \text{for individual } i \text{ with } AA \text{ genotype} \\ \text{Bin}(X_i, 0.5) & \text{for individual } i \text{ with } AB \text{ genotype} \\ 0 & \text{for individual } i \text{ with } BB \text{ genotype} \end{cases}$$

Once DNA samples from  $N$  individuals are combined into a pool, the total number of molecules in the pool can be denoted  $X$  ( $X = \sum_{i=1, \dots, N} X_i$ ) and the total number of A alleles in the pool can be denoted  $Y$  ( $Y = \sum_{i=1, \dots, N} Y_i$ ). Then the value of  $p''$  (true allele frequency in the pool) is the ratio  $Y/X$ . If the pool is assayed on a single chip, the allele frequency estimate (estimate of  $p$ ) obtained from the chip would be

$$r = p'' + W,$$

where  $W$  is additional measurement error introduced by reading the chip. We model this as normally distributed with mean 0 and variance  $\sigma_w^2$ . Putting all levels of this model together, the mean and variance of the allele frequency estimate,  $r$ , would be as follows.

$$E(r) \approx p$$

and

$$\text{Var}(r) \approx \text{Var}(p'/p) + \text{Var}(p''/p') + \text{Var}(W) \approx \frac{p(1-p)}{2N} + \frac{p'(1-p')}{2N} \tau^2 + \sigma_w^2.$$

The first term,  $Var(p'/p)$ , represents sampling from the population. The second term,  $Var(p''/p')$ , is the effect of pooling the individual samples. The third term,  $Var(W)$ , is pure measurement error. The expression for  $Var(p''/p')$  can be derived as  $\frac{p'(1-p')}{2N} \tau^2$  in Jawaid et al. (2002). If we are willing to use the approximation  $p' \approx p$ , then the expression for  $Var(r)$ , can be simplified as follows.

$$Var(r) \approx \frac{p(1-p)}{2N} (1 + \tau^2) + \sigma_w^2.$$

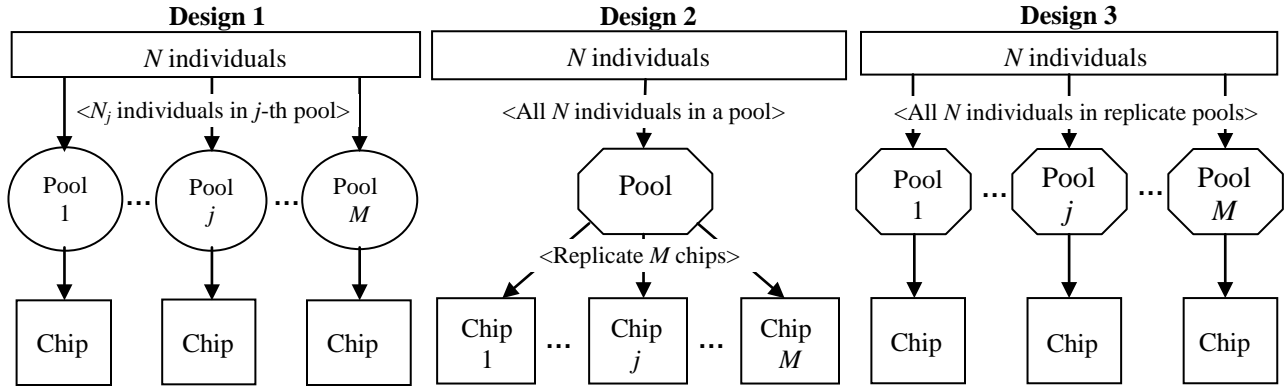
In addition, if we assume that  $\tau$  is very small (as discussed above), with value with typically less than 0.1, then the term containing  $\tau^2$  will be close to zero. Computations of the mean and variance of  $r$ ,  $Var(r)$ , are in Appendix A.

### 3.3 DESIGNS

#### 3.3.1 Three Designs for a Case-Control Study

We consider three standard simple designs case-control designs using pooled DNA. To directly compare efficiency, we consider designs with the same number of chips and same number of study subjects. At this point we assume that there are no covariates (e.g. sex) that need to be accounted for in the design; covariates are considered in the discussion. Let  $N$  be the number of individuals in each cohort (case and control), and let  $M$  be the number of of chips for each cohort. Design 1 divides the  $N$  individuals into  $M$  groups, so the  $j$ th group ( $j = 1, \dots, M$ ) consists of  $N_j$  individuals ( $N = \sum_j N_j$ ). Then we generate one pool from each of the  $M$  groups and assay each

pool on one chip (Figure 3.2). For design 2, we generate one pool with all  $N$  individuals and assay that pool on  $M$  replicate chips. Design 3 is similar to design 2 except that we generate the pool of all  $N$  individuals  $M$  independent times, assaying each pool on a chip.



**Figure 3.2: Three designs for non-covariate model**

### 3.3.2 Allele Frequency Estimates from Each Design

Building on the model introduced above, for case-control study let the true allele frequency of the cohort be  $p_g$  ( $g = 1$  for case, 2 for control). For each of the designs discussed above, we then have slightly different notation for the allele frequencies in the pool, as shown in Table 3.1, but the sample size are all the same – total number of individuals and total number of chips. Then the allele frequencies for the  $j$ th chip in the  $g$ th cohort group,  $r_{gj}$  ( $g = 1$  for case, 2 for control;  $j = 1, \dots, M$ ), for the three designs would be

$$\text{Design 1 and 3:} \quad r_{gj} = p''_{gj} + W_{gj} \quad (1)$$

$$\text{Design 2:} \quad r_{gj} = p''_g + W_{gj} \quad (2)$$

**Table 3.1: True allele frequencies of three designs**

True Allele Frequency	Design 1	Design 2	Design 3
Population	$p_g$	$p_g$	$p_g$
Samples from the population	$p'_g$	$p'_g$	$p'_{gj}$
One pool from samples	$p''_{gj}$	$p''_g$	$p''_{gj}$
One Chip	$r_{gj}$	$r_{gj}$	$r_{gj}$

Notation:  $g$  = indicator of cohort group; and  $j$  = indicator of pools/chips ( $j = 1, \dots, M$ )

### 3.4 TEST STATISTICS

We consider three possible ways to test the null hypothesis that there is no difference in allele frequencies between case and control groups: a standard t-test, a modified t-test, and a chi-squared test. If we treat each chip as a single observation and let  $r_{gj}$  ( $g = 1$  for case, 2 for control;  $j = 1, \dots, N$ ) be the  $N$  observations from two independent groups (case and control) with sample means,  $\bar{r}_g$ , then the standard two-sample t-test statistic would be

$$T = \frac{\bar{r}_{1.} - \bar{r}_{2.}}{\sqrt{\frac{Var(\bar{r}_{1.}) + Var(\bar{r}_{2.})}{2N}}}.$$

This would be treated as following a  $t$ -distribution, and the variance in the denominator of  $T$  would be calculated by the usual pooled variance estimate. Since, however, pooled DNA data has a mixture binomial and normal variation (modeled as above), we expect that this standard two-sample t-test would be anti-conservative for pooled data. Therefore we generate a new test

statistic (called *Modified T-test*), whose statistic formula looks like the standard two-sample t-test, with adjustment for the additional binomial variance. The variances of the allele frequency estimates consist of three terms (See Chapter 3.2 and Appendix B.1). The first two terms are for binomial variation introduced by sampling individuals and then DNA strands, and the last term is for Gaussian variation introduced by reading the chip. However, we assume that the second term is likely to be very small because of tiny  $\tau$ . If we are not willing to make this assumption, then our variance formula requires an estimate of  $\tau$  in order to apply it to real data. Then our modified t-test formula would be

$$T = \frac{\bar{r}_1 - \bar{r}_2}{\sqrt{Var(\bar{r}_1) + Var(\bar{r}_2)}} .$$

The third test that we consider is the standard  $\chi^2$ -test based on the 2×2 contingency table of two alleles (Locus A/Locus B) and cohort group (case/control). This table is not observed directly, but must be inferred (estimated) from the allele frequency estimates and the total sample size. There are  $N$  individuals in each cohort (case/control). Then the total numbers of alleles for case and control are fixed,  $2N$  each. The number of each allele in each group would be calculated by multiplying the allele frequency of the allele by  $2N$ . That is,  $N_{A1} = p_{1.} \times 2N$  and  $N_{B1} = (1 - p_{1.}) \times 2N = 2N - N_{A1}$  for cases, and  $N_{A2} = p_{2.} \times 2N$  and  $N_{B2} = (1 - p_{2.}) \times 2N = 2N - N_{B1}$  for controls, where  $p_{g.}$  is the allele frequency of locus A for  $g$ th cohort group. When the standard chi-squared test is applied to this contingency table the result is again clearly anti-conservative, since we have not accounted for the error in observation of the numbers of alleles.

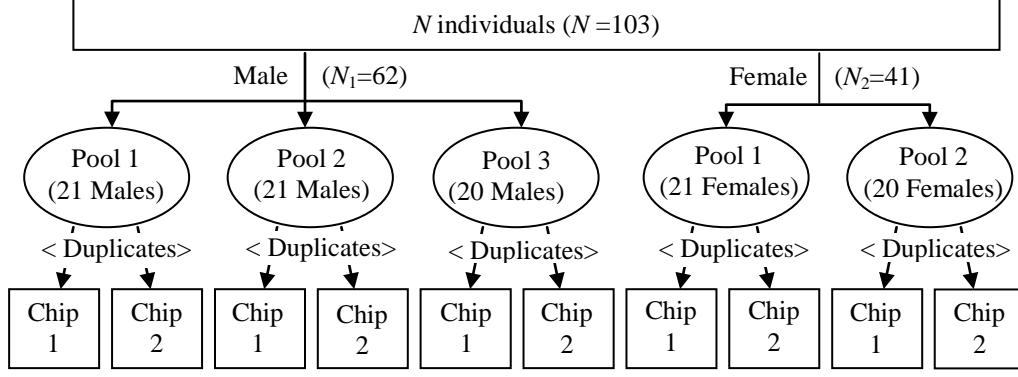
**Table 3.2: Allele frequency table for  $\chi^2$ -test**

	Case	Control
Locus $A$	$N_{A1}$	$N_{A2}$
Locus $B$	$N_{B1}$	$N_{B2}$
Total	$2N$	$2N$

Since both the standard two-sample t-test and the  $\chi^2$ -test do not consider the pooling strategy, they can be applied to any of the pooling design we considered without modification. The modified t-test has a slightly different denominator depending on the pooling design. We can compute modified t-test statistics for the designs we suggested, and we can also find out which design is most efficient by comparing the modified t-test statistics (equivalently, comparing the variances of allele frequency estimates) of the designs.

### **3.5 PANCREATIC CANCER AFFYMETRIX 6.0 POOLED DATA**

Diergaarde et al. (in press) describes a case-control study for pancreatic cancer in which the first stage of genotyping was performed in pools using the Affymetrix 6.0 chip. There were 103 cases and 103 controls with 906,600 SNPs, and each cohort was divided into five pools of approximately 20 people. Each pool was replicated twice on two chips. This is a hybrid of the designs discussed above, but is easily analyzed by extending the statistics discussed above to an ANOVA framework.



**Figure 3.3: Pooled affymetrix 6.0 pancreatic cancer data**

## 3.6 RESULTS

### 3.6.1 Relative Efficiency of the Three Designs

The allele frequency estimate from each chip consists of two terms as shown in (1) and (2) in Chapter 3.3.2. The first term has binomial variation as a result of both population sampling and pooling, although we have argued that the pooling component is typically negligible, and the second term is Gaussian variation introduced by the process of reading the chip. The variances of the allele frequencies,  $Var(r_{gj})$ , for each design are as follows (See Appendix B.1).

$$\text{Design 1:} \quad Var(r_{gj}) \approx \frac{p_g(1-p_g)}{2N_j} + \frac{p'_{gj}(1-p'_{gj})}{2N_j} \tau^2 + \sigma_w^2$$

$$\text{Design 2 and 3:} \quad Var(r_{gj}) \approx \frac{p_g(1-p_g)}{2N} + \frac{p'_g(1-p'_g)}{2N} \tau^2 + \sigma_w^2$$



If we assume that  $p'_g \approx p_g$ , the variances of the overall allele frequency estimates for each design can be written as follows.

$$\text{Design 1: } \text{Var}(\bar{r}_{g.}) \approx \frac{1}{M^2} \sum_j \left[ \frac{p_g(1-p_g)}{2N_j} + \frac{p_g(1-p_g)}{2N_j} \tau^2 \right] + \frac{\sigma_w^2}{M} \quad (3)$$

$$\text{Design 2: } \text{Var}(\bar{r}_{g.}) \approx \frac{p_g(1-p_g)}{2N} + \frac{p_g(1-p_g)}{2N} \tau^2 + \frac{\sigma_w^2}{M} \quad (4)$$

$$\text{Design 3: } \text{Var}(\bar{r}_{g.}) \approx \frac{p_g(1-p_g)}{2N} + \frac{p_g(1-p_g)}{2NM} \tau^2 + \frac{\sigma_w^2}{M} \quad (5)$$

The variance of Design 1 in (3) is identical to the variance of Design 2 in (4). Thus Design 1 and 2 are equivalent from a statistical efficiency point of view. Design 3 is superior (lower variance). This improvement in efficiency is attributable to the fact that in design 3 each DNA sample is measured and pooled repeatedly, whereas in designs 1 and 2 each sample is only measured and pooled once. The difference in the variances between designs, however, is only in the second term, which involves the  $\tau^2$  factor. Our work and that of others has suggested that in carefully-performed pooling studies  $\tau^2$  is generally quite small. If that is true, then all three designs have approximately the same variance and thus are approximately equally efficient.

### 3.6.2 Comparison of the Three Test Statistics using the Affymetrix 6.0 Pooled Data

To test the null hypothesis that there is no difference in allele frequencies between cases and controls, we consider the standard two-sample t-test, modified t-test, and  $\chi^2$ -test as discussed

above. There are two things to be careful in this dataset before applying test statistics for comparisons. First, there is a covariate (gender effect) in this design. However our models are for non-covariate designs. So we just ignored the gender effect for now to compare test statistics. Second, this data design is not a perfect match with any of our designs even if a covariate effect (gender) is ignored. Since the design can be considered as a mixture form of our designs 1 and 3 (in Figure 3.2), it is hard to apply our modified t-test statistic directly. Therefore we considered two ways to apply modified t-test statistics on this data. As mentioned above, the gender effect is not considered for now. There are duplicate chips from a pool in this data. Then the design of each duplicate is the same as Design 1 in Figure 3.2, since each single chip comes from single subpool for each duplicate. So we applied test statistics to each duplicate and called the experiments “*Replicate 1*” and “*Replicate 2*,” respectively. Alternatively we can consider all duplicate chips at the same time; however we need to modify the variance of overall allele frequency estimate for the modified t-test statistic beforehand. This analysis is denoted as “*Both Replicates Combined*.”

We compared the three statistics for each dataset (Replicate 1, Replicate 2, and Both replicates combined) by the following procedure. We first applied the standard two-sample t-test to the data. Among all 906,600 SNPs, 71,648 SNPs (7.9%) had significant p-values for the standard two-sample t-test at significance level  $\alpha = 0.05$ . We then ranked SNPs based on the t-test statistics and chose 200 SNPs (100 SNPs from the top and 100 from the bottom) with the largest test statistics. All p-values of the 200 selected SNPs were very small (maximum p-value of the 200 SNPs was 0.000134). Next we applied our modified t-test and the  $\chi^2$ -test to these 200 SNPs. Table 3.3 shows the numbers of significant SNPs among the selected 200 SNPs at significance level  $\alpha$  is 0.05. “*Replicate 1*” and “*Replicate 2*” refer to the two replicate datasets,

and “*Both Replicates Combined*” means to consider all duplicates simultaneously, as described above.

**Table 3.3: Number of significant SNPs among selected 200 SNPs of modified t-test and  $\chi^2$ -test at  $\alpha = 0.05$**

	Modified $T$ -test	$\chi^2$ -test
Replicate 1	30/200	65/200
Replicate 2	39/200	70/200
Both Replicates Combined	55/200	73/200

For both the modified t-test and the  $\chi^2$ -test not all 200 SNPs have p-values less than 0.05, even though all 200 SNPs have very tiny p-values by the standard two-sample t-test. Thus we infer that the standard two-sample t-test is extremely anti-conservative as compared to the other tests (modified t-test and  $\chi^2$ -test). From Table 3.3, it is also evident that the  $\chi^2$ -test is more liberal than the modified t-test. The number of significant SNPs from the  $\chi^2$ -test is larger than from the modified t-test, and in fact all SNPs found significant by the modified t-test are also detected by the  $\chi^2$ -test. Therefore, as expected, the modified t-test is the most conservative among the three test statistics. It is interesting to note, however, that the difference between the modified t-test and the chi-squared test is relatively modest compared to the extremely anti-conservative behavior of the standard t-test. From this we can conclude that the Gaussian portion of the variability in our pooling model is small relative to the binomial sampling variability, and so it is most essential for any test statistic to account for the binomial variability.

Since the modified t-test accounts for both the binomial and Gaussian components of variability, it in theory has correct type I error and should be the most appropriate test statistic. However, there is one further consideration. In a typical pooling experiment, both the standard and modified t-tests are computing the Gaussian component of the variance from a very small number of chips. This type of statistic, applied in a genome-scan setting, often has unacceptable performance because the SNPs that have the smallest p-values are those that have the lowest estimated Gaussian variance components by random chance. This problem has been discussed at length in the expression microarray literature (Lin et al., 2008), and usual solution has been a shrinkage estimator. Such an approach could be applied in the current problem, but we suggest that if the primary goal is stable ranking of SNPs for follow-up then the simple chi-squared test might be equally appropriate. Our results above show that it is anti-conservative, but not by too much, and it should have much more stable ranking behavior than any statistic that attempts to incorporate Gaussian variability based on small-sample estimates.

### **3.6.3 What is the Best Pooling Design?**

We suggested three commonly-used pooling designs with the same sample (chip) size and the same number of individuals to investigate the most effective design strategy. We calculated the variances of the allele frequency estimates from each design in order to compare the efficiency of the designs. The variances of designs 1 and 2 are the same, whereas the variance of design 3 is theoretically less than that of designs 1 and 2. However, in practice there is probably not a large difference because we expect the contribution of pooling variability to be small. If there is concern about the pooling variability, then clearly design 3 is preferable, but otherwise there is

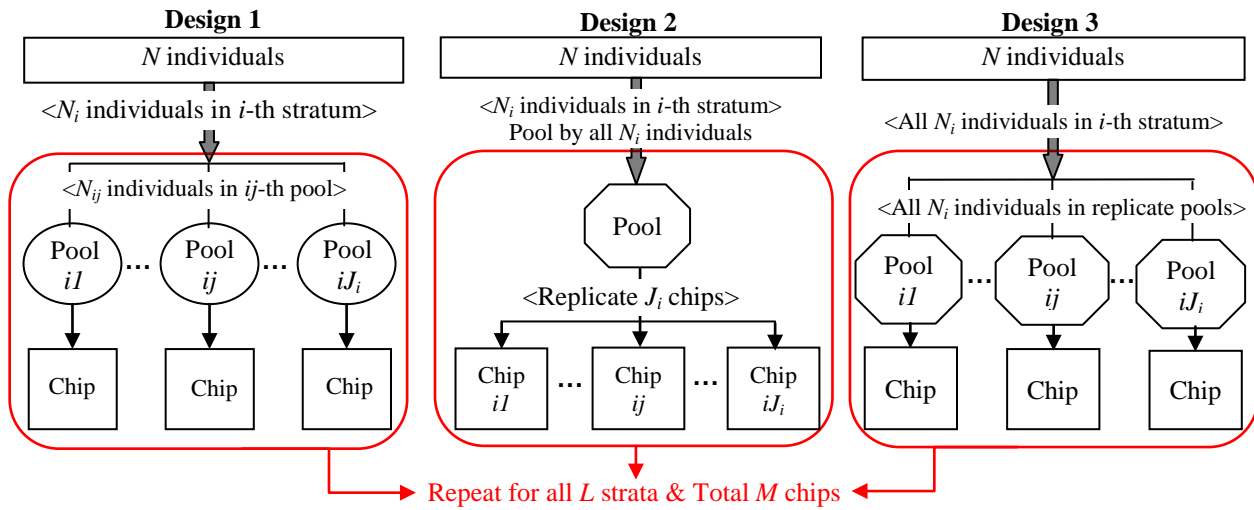
no need to do the extra labor of repeatedly quantifying each sample, and design 1 or design 2 should suffice. Although designs 1 and 2 are statistically equivalent under our model, there might be a slight practical advantage to design 1 - it might be somewhat more robust to potential lab errors, for example in the handling of the pooled DNA.

### **3.6.4 What is the most Appropriate Test Statistic?**

We derived a modified t-test statistic under our pooling model for each design, which also extends easily to more complex ANOVA-type designs. The modified t-test incorporates both binomial sampling variability and Gaussian variability into the denominator. We then compared the modified t-test to the standard t-test and a chi-squared test based on a reconstructed allele count table. Our modified t-test is in theory the most appropriate statistic, since it correctly models the variance of the allele frequency difference. We applied the three test statistics to a dataset of Affymetrix 6.0 pooled data. Our results showed that the standard two-sample t-test is very strongly liberal compared to the modified t-test and  $\chi^2$ -test. The chi-squared test is slightly more liberal than the modified t-test, but much less so than the standard t-test. We conclude that the standard t-test is clearly inappropriate for this type of analysis, but that both the modified t-test and the chi-squared test might be reasonable alternatives. Based on previous results in the expression microarray literature, we suggest that the  $\chi^2$  statistic might have acceptable Type I error and have more stable behavior for ranking genes than the modified t-test. Further work to test this hypothesis might include genome-wide simulations to see how ranked lists of genes (as opposed to just p-values of individual genes) vary between the statistics.

### 3.6.5 Designs with Covariates

All of our analyses above assumed that the pooling design does not need to incorporate covariates. However, it is not unusual to include covariates such as sex in pooled studies. The three designs we suggested in this study are also applicable for models with covariates, if each design is repeated in each stratum of the covariate. The alternative is to make each pool mixed (e.g. combined male and female). (See Figure 3.4). Advantages and disadvantages of these approaches should be considered in future work, as well as appropriate test statistics for each.



**Figure 3.4: Three designs of covariate model**

Intuitively the results of the three covariate designs are the same as for the non-covariate designs. In the pooling strategy for the covariate model, one issue we have to consider is how to pool individuals when covariates exist. For instance, suppose there is one covariate, SEX. There are two ways for pooling. One is to pool individuals stratified by SEX – pools only for males and females respectively. The other is to mix the two sexes and to put them in one pool – for instance,

mix sexes 50 to 50 in each pool. Which pooling strategy is better for detecting genetic effects? In the further study, we will find out the most effective pooling strategy with covariates.

## **4.0 CONCLUSION AND DISCUSSION**

This dissertation discusses improved methods for analyzing raw data from genotyping assays, with particular attention to two specific problems. The first is calling genotypes in individuals with non-standard numbers of chromosomes (e.g. trisomy), and the second is testing genotype-phenotype associations using pooled genotype data.

### **4.1 NEW GENOTYPE CALLING METHOD FOR TRISOMIC INDIVIDUALS**

We suggest a modified algorithm to call the genotypes of trisomic individuals with Down syndrome, who have an extra copy of chromosome 21. Many clustering algorithms to call genotypes have been developed and used for disomic individuals. However when a chromosomal abnormality has occurred, the regular algorithms do not work properly. Lin et al. (2008) suggested a parametric mixture algorithm to call the genotypes of Down syndrome individuals. However it sometimes miscalled ambiguous genotypes between the clusters, when compared to Illumina calls of the SNP. Our algorithm is an updated version of Lin's algorithm with a little different assumption. Under the assumption that Illumina calls are correct and homozygote Illumina calls are perfect, we apply a new parametric mixture model (called HINM) to heterozygote individuals only. The genotype calling result depends on the quality of the SNP.



Based on the genotype calling results from two real datasets, HINM method looks less sensitive than Lin's IBM method to the data quality.

Both Lin's IBM and our HINM methods called the genotypes pretty well, but the results are not always the same. Theoretically trisomic individuals should have four possible parallel genotype clusters, but some SNPs do not. For high-quality SNPs with very clean distinct clusters, both methods work well and there is no difference between the results of the two algorithms. Therefore the comparison is performed based on scatter plots and call results using mismatch rate and SNP quality for SNPs with at least one mismatched call. Overall HINM method is better for many SNPs regardless of mismatch rate and SNP quality. In the case of SNPs with low-intensity curvature, usually the HINM method runs pretty well. Even though SNPs have muddy clusters, HINM classify the clusters better, but for some of them Lin's method is better but not much. In most cases of SNPs with better calls by Lin's method, there are suspicious Illumina calls at the first such as one homozygote Illuminas are missing or one homozygote (or heterozygote) Illuminas have more than one clusters, even though it is ideal that there are four clusters (two for homozygote; two for heterozygote) in Illumina call plots.

Our method has some limitations that should be mentioned. We applied our method to Illumina data. For data from other platforms, HINM could be also useable after transforming the raw data to 1-dimensional data, similarly to Lin's IBM method. And HINM method is suggested only for trisomic individuals like Lin's IBM method. However it is possible that various chromosomal abnormality occurred. Usually we can get 2-dimensional raw data from the genotype assays. However it is much harder to get more precise results for more complex chromosomal abnormality using the 2-dimensional data. The more complicated the chromosomal abnormality that exists, the more information from the raw data is needed. Therefore in the

further study it is one of the big issues whether the genotype calling methods for trisomic individuals can be used for more severe chromosomal abnormality. And there is no standard definition for good and bad calls. Other standards could be possible to classify the call results into good or bad calls. However in any case the performance for SNPs with bad calls is not as important as the performance for SNPs with good calls.

## **4.2 MOST EFFICIENT POOLING STRATEGES AND TEST STATISTICS**

Pooled DNA genotyping is used as a pre-screening method in a genetic association study because of the high cost of individual genotyping of thousands of individuals. After the most likely candidate genes based on pooling genotyping, most pooling studies do individual genotyping for the candidates to test for the genes related to disease.

An efficient pooling strategy is important to get more accurate results. We considered three different pooling designs with the same number of individual samples and chips. Theoretically the design called Design 3 in Chapter 3 is the best pooling design, which has multiple replicate pools using all individual samples.

We also considered test statistics for pooled genotyping. To find the most appropriate test statistics, we considered a modified t-test statistic that includes both binomial variation and normal variation simultaneously due to pooling. Then we compared three tests statistics – the standard two-sample t-test, the modified t-test we suggested, and a  $\chi^2$ -test for contingency tables. Among three test statistics, the modified t-test is the most appropriate test for pooled DNA data theoretically. We applied the three test statistics to a dataset of Affymetrix 6.0 pooled data to compare them. The standard two-sample t-test is the most liberal statistic, while the modified t-

test is the most conservative statistic. The results by  $\chi^2$ -test are not too different from the modified t-test, however, even though the  $\chi^2$ -test only considers binomial sampling effect.

In further study we need to consider covariates. We considered the simplest design in this study, in which any covariates are not considered. When covariates are included in the model, the design strategies in this study can be extended to the designs for a covariate model. One more thing we need to consider is how to pool when an important covariate exists just like mentioned in Chapter 3.6.5.

### **4.3 MORE GENERAL COMMENTS ON USES FOR RAW GENOTYPE DATA AND RELATED PROBLEMS**

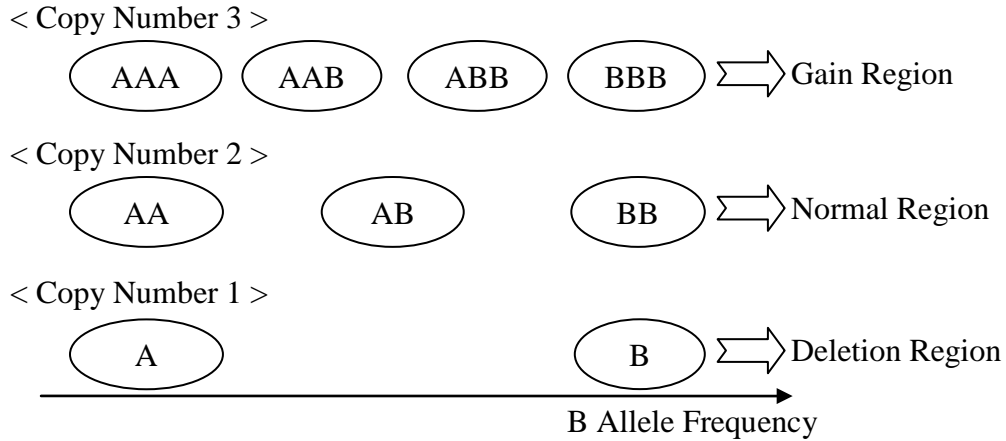
In addition to the two issues we considered, there are many other statistical issues related to use of raw genotype assay data. In the context of genotype calling, some of the other issues include SNP data quality, low-intensity points, and CNV calling.

Many clustering algorithms usually work well for disomic individuals. And SNP quality is important for genotype calling procedures. For low-quality SNPs, clustering method miscalled the genotypes sometimes. Therefore SNP quality measures like RSS are used for disomic individuals in Affymetrix. However the clustering algorithm is more complex when chromosomal abnormality exists. Some modifications are needed for clustering methods when there is chromosomal abnormality. Our genotype calling method is modified for trisomic individuals. However we need more careful considerations for other chromosomal abnormalities like tetrasomic individuals. Moreover we scored SNP qualities by eye in our study. In the future, it is better to consider possible methods for SNP quality classification automatically. After calling,

there is lack of numerical measures to check the calling results. We inspected the plots by eye to figure out whether there were good calling results or not. However if there are lots of SNPs in the dataset, this method is time consuming.

Since SNPs have genotype groups (i.e., three groups for disomic individuals, and four groups for trisomic individuals) with different slopes each in 2-dimensional intensity plots, transformation for 1-dimension is usually used to distinguish them easily. However some SNPs have low-intensity curvatures after transformation. Usual clustering methods failed to classify them clearly. In the case of trisomic individuals, four genotype clusters exist - two upper groups and two lower groups are tending to regress into the middle of two heterozygote. It is a concern how to call individuals with low-intensity curvature clearly. We might need a specific cut-off point to define the low-intensity curvature. It is another issue how to choose the cut-off point. In addition, it is worth considering other transformation formulae in order to classify groups more appropriately.

There are also uses of raw genotype assay data that are very different from those we considered. One issue is the study of copy number variation (CNV). CNV is duplication or deletion of segments of genome compared to a reference genome. Genotype assays from Affymetrix or Illumina are high-throughput arrays, and these arrays contain millions markers. In addition, high-throughput arrays are designed for genotyping and are allele-specific. CNVs are usually called by looking at the total intensity and looking for several SNPs in a row where the intensity is low or high (Figure 4.1). You and Holmes (2008) provided the overview of various computational approaches to CNV discovery using SNP genotyping. If we have a SNP where we know there is a CNV, we ought to be able to call the genotype and copy number by looking at a



**Figure 4.1: Generalized genotyping**

single scatter plot and using clustering algorithms. It might be useful to get much bigger picture than a single SNP and it might be possible to detect more specific characteristics. Usually change-point method and hidden markov model are used for CNV calling. And the chromosome-wide shape for genotypes is shown by B-allele frequencies in a whole chromosome. To use our method for CNV calling, we need to consider how to expand and apply our method for CNV. The first step of CNV calling is regularly starting with SNP genotype calling to verify unusual patterns of genotypes, and then segmentation approaches are used to find CNVs after genotype calling. For now, however, the popularly used genotype calling algorithm is not settled down to call genotypes for trisomic or upward individuals. Our method is developed for single SNP marker at a time for trisomic individuals. Chromosomal-wide genotyping using SNP arrays can induce to discover CNVs. Therefore we can expand our method for mining genotypes in whole genome. One thing to be careful is that our method is developed for only trisomic individuals, not parent-offspring trios. Normally Mendelian inconsistency is revealed from parent-offspring

trios. It might be difficult to detect Mendelian inconsistency from only using trisomic or more individuals.

## APPENDIX A

### MEAN AND VARIANCE OF THE ALLELE FREQUENCY

Based on the definition of  $X_i$  and  $Y_i$  in Chapter 3.2, computations of the second term,  $Var(p''/p')$ , are as follows (similar with Jawaaid et al., 2002).

$$\begin{aligned}
 E(X) &= n\mu & Var(X) &= n\tau^2\mu^2 \\
 E(Y) &= N_{AA}\mu + N_{AB}\mu/2 = \mu p' n & Var(Y) &= N_{AA}\tau^2\mu^2 + N_{AB}(\mu/4 + \tau^2\mu^2/4) \\
 Cov(X, Y) &= \tau^2\mu^2 p' n
 \end{aligned}$$

Using the formula in Mood, Graybill, and Boes (1974), which is approximate formulas for mean and variance of the quotient of two correlated variables, the mean and the variance of  $r$  are calculated as follows.

1) *Mean of the allele frequency,  $E(r)$ :*

$$\begin{aligned}
 E(r) &= E(p'') + E(W) \\
 &= E[E(p''/p')] && \text{with } E(W) = 0 \\
 &\approx E(p') && \text{where } E(p''/p') = E\left(\frac{Y}{X} \mid p'\right) \approx p' \text{ by the formula} \\
 &= p
 \end{aligned}$$

2) *Variance of the allele frequency,  $Var(r)$ :*

$$\begin{aligned}
 Var(r) &= Var(p'') + Var(W) \\
 &= Var[E(p''/p')] + E[Var(p''/p')] + Var(W) \\
 &= Var(p'/p) + Var(p''/p') + Var(W)
 \end{aligned}$$

The first and third terms,  $Var(p'/p)$  and  $Var(W)$ , are already known as  $\frac{p(1-p)}{2N}$  and  $\sigma_w^2$ ,

respectively. However the calculation of second term,  $Var(p''/p')$ , is not simple. Then, by the formula,  $Var(p''/p')$  is computed approximately as follows.

$$Var(p''/p') = Var\left(\frac{Y}{X} / p'\right) \approx \frac{1}{N} \left[ \frac{N_{AB}/N}{4\mu} + \tau^2 \left( N_{AA}/N + \frac{N_{AB}/N}{4} - p'^2 \right) \right]$$

If we assume large  $\mu$  and HWE, the second term,  $Var(p''/p')$ , approximately would be

$$Var(p''/p') \approx \frac{p'(1-p')}{2N} \tau^2$$

where  $p' = (2N_{AA} + N_{AB})/2N$ . Therefore the variance of  $r$  would be

$$Var(r) \approx \frac{p(1-p)}{2N} + \frac{p'(1-p')}{2N} \tau^2 + \sigma_w^2.$$



## APPENDIX B

### VARIANCE OF THE ALLELE FREQUENCY FOR TEST STATISTICS

#### B.1 VARIANCE OF THE ALLELE FREQUENCY FOR THREE DESIGNS

*Design 1:*

The variance of allele frequency of the  $j$ th chip,  $r_{gj}$  would be

$$Var(r_{gj}) = Var(p''_{gj}) + Var(W_{gj}).$$

The first term,  $Var(p''_{gj})$ , is the variance of the allele frequency of the pool, which  $j$ -th chip comes from. This contains the sampling variation including the pooling variation. The sampling variation from the population would be common over the chips, whereas the pooling variation would be vary depending on the pool, from which the chip comes. These two variations are not independent each other and have the binomial effects, but the pooling variation would be very tiny compared to the sampling variation.

$$Var(p''_{gj}) = Var[E(p''_{gj}|p'_{gj})] + E[Var(p''_{gj}|p'_{gj})]$$

The first term,  $Var[E(p''_{gj}|p'_{gj})]$ , is about the sampling variation and would be the similar with the variance of the allele frequency of subindividuals. And the second term,  $E[Var(p''_{gj}|p'_{gj})]$  is

the variation due to pooling. Each pool consists  $N_j$  individuals with allele frequency  $p'_{gj}$ .

Therefore the variance of the allele frequency of the pool,  $Var(p''_{gj})$ , would be as followed.

$$\begin{aligned} Var(p''_{gj}) &= Var(p'_{gj}/p_g) + E \left[ \frac{p'_{gj}(1-p'_{gj})}{2N_j} \tau^2 \right] \\ &\approx \frac{p_g(1-p_g)}{2N_j} + \frac{p'_{gj}(1-p'_{gj})}{2N_j} \tau^2. \end{aligned}$$

where  $\tau$  is the coefficient of variation of the number of DNA molecules of locus  $A$  contributed by each individual. Then the variance of the allele frequency of each chip approximately would be

$$Var(r_{gj}) = \frac{p_g(1-p_g)}{2N_j} + \frac{p'_{gj}(1-p'_{gj})}{2N_j} \tau^2 + \sigma_w^2.$$

The variance of the overall allele frequency would be

$$Var(\bar{r}_g) = \frac{1}{M^2} \sum_j \left[ \frac{p_g(1-p_g)}{2N_j} + \frac{p'_{gj}(1-p'_{gj})}{2N_j} \tau^2 \right] + \frac{\sigma_w^2}{M^2}.$$

Similarly, the variances of Design 2 and Design 3 could be calculated.

*Design 2:*

The variance of allele frequency of the  $j$ -th chip,  $r_{gj}$  would be

$$Var(r_{gj}) = Var(p''_g) + Var(W_{gj}).$$

The first term,  $Var(p''_g)$ , is the variance of the allele frequency of the pool.

$$Var(p''_g) = Var \left[ E(p''_g | p'_g) \right] + E \left[ Var(p''_g | p'_g) \right]$$

The first term,  $Var \left[ E(p''_g | p'_g) \right]$ , would be the same as the variance of the allele frequency of individuals,  $Var(p'_g)$ . And the second term,  $E \left[ Var(p''_g | p'_g) \right]$  is the variation due to pooling. All

$N$  individuals are in the pool with allele frequency  $p'_g$ . Therefore the variance of allele frequency of the pool,  $Var(p''_g)$ , would be

$$\begin{aligned} Var(p''_g) &= Var(p'_g/p_g) + E \left[ \frac{p'_g(1-p'_g)}{2N} \tau^2 \right] \\ &\approx \frac{p_g(1-p_g)}{2N} + \frac{p'_g(1-p'_g)}{2N} \tau^2. \end{aligned}$$

The variance of the allele frequency of each chip would be

$$Var(r_{gj}) \approx \frac{p_g(1-p_g)}{2N} + \frac{p'_g(1-p'_g)}{2N} \tau^2 + \sigma_w^2.$$

Then the variance of the overall allele frequency would be

$$Var(\bar{r}_g) = \frac{p_g(1-p_g)}{2N} + \frac{p'_g(1-p'_g)}{2N} \tau^2 + \frac{\sigma_w^2}{M^2}$$

*Design 3:*

The variance of allele frequency of the  $j$ th chip,  $r_{gj}$  would be

$$Var(r_{gj}) = Var(p''_{gj}) + Var(W_{gj}).$$

The first term,  $Var(p''_{gj})$ , is the variance of the allele frequency of the pool.

$$Var(p''_{gj}) = Var \left[ E(p''_{gj}|p'_g) \right] + E \left[ Var(p''_{gj}|p'_g) \right]$$

The first term,  $Var \left[ E(p''_{gj}|p'_g) \right]$ , would be the same as the variance of the allele frequency of individuals,  $Var(p'_g)$ . And the second term,  $E \left[ Var(p''_{gj}|p'_g) \right]$  is the variation due to pooling. All  $N$  individuals are in each pool with allele frequency  $p'_g$ . Therefore the variance of allele frequency of the pool,  $Var(p''_{gj})$ , would be

$$Var(p''_{gj}) = Var(p'_g/p_g) + E \left[ \frac{p'_g(1-p'_g)}{2N} \tau^2 \right]$$

$$\approx \frac{p_g(1-p_g)}{2N} + \frac{p'_g(1-p'_g)}{2N} \tau^2$$

Then the variance of the allele frequency of each chip would be

$$Var(r_{gj}) \approx \frac{p_g(1-p_g)}{2N} + \frac{p'_g(1-p'_g)}{2N} \tau^2 + \sigma_w^2.$$

Thus the variance of the overall allele frequency would be

$$Var(\bar{r}_{g.}) = \frac{p_g(1-p_g)}{2N} + \frac{p'_g(1-p'_g)}{2NM} \tau^2 + \frac{\sigma_w^2}{M^2}.$$

## B.2 PREDICTED VARIANCE OF THE OVERALL ALLELE FREQUENCY

Since we don't know the true allele frequency of each cohort,  $p_g$ , we can recalculate the variances using estimated allele frequencies of each chips,  $r_{gj}$ 's. So  $\widehat{Var}(\bar{r}_{g.})$  would be

$$\text{Design 1 and 2: } \widehat{Var}(\bar{r}_{g.}) = \frac{\bar{r}_{g.}(1-\bar{r}_{g.})}{2N} + \frac{\bar{r}_{g.}(1-\bar{r}_{g.})}{2N} \tau^2 + \frac{\sum_j (r_{gj} - \bar{r}_{g.})^2}{M(M-1)}$$

$$\text{Design 3: } \widehat{Var}(\bar{r}_{g.}) = \frac{\bar{r}_{g.}(1-\bar{r}_{g.})}{2N} + \frac{\bar{r}_{g.}(1-\bar{r}_{g.})}{2NM} \tau^2 + \frac{\sum_j (r_{gj} - \bar{r}_{g.})^2}{M(M-1)}$$

There is one restriction in the variances. If there are a lot of samples we have, the variances could be used as well. If not, however, the third term of the variances,  $\frac{\sum_j (r_{gj} - \bar{r}_{g.})^2}{M(M-1)}$ , could not be used for the variances, because this term came from under the assumption of the normality. In this case, other way to get the variances has to be used.

## BIBLIOGRAPHY

- Bader, J. S., Bansal, A., and Sham, P. (2001). Efficient SNP-based tests of association for quantitative phenotypes using pooled DNA. *GeneScreen*, 1:143-150.
- Bader, J. S. and Sham, P. (2002). Family-based association tests for quantitative traits using pooled DNA. *Eur J Hum Genet*, 10:870-878.
- Barnes, M., Freudenberg, J., Thompson, S., Aronow, B., and Pavlidis, P. (2005). Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucl. Acids Res.*, 33:5914-5923.
- Craig, D. W., Huentelman, M. J., Hu-Lince, D., Zismann, V. L., Kruer, M. C., Lee, A. M., Puffenberger, E. G., Pearson, J. M., and Stephan, D. A. (2005). Identification of disease causing loci using an array-based genotyping approach on pooled DNA. *BMC Genomics*, 6:138.
- Dalma-Weiszhausz, D. D., Warrington, J., Tanimoto, E. Y., and Miyada, C. G. (2006). The affymetrix GeneChip platform: an overview. *Methods Enzymol*, 410:3-28.
- Di, X., Matsuzaki, H., Webster, T. A., Hubbell, E., Liu, G., Dong, S., Bartell, D., Huang, J., Chiles, R., Yang, G., Shen, M.-m., Kulp, D., Kennedy, G. C., Mei, R., Jones, K. W., and Cawley, S. (2005). Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics*, 21:1958-1963.
- Diergaarde, B., Brand, R., Cheong, S. Y., Lamb, J., Stello, K., Barmada, M. M., Feingold, E., and Whitcomb, D. C. ((in press)). Pooling-based genome-wide association study implicates GGT1 (gamma-glutamyltransferase 1) in pancreatic carcinogenesis. *Pancreatology*.
- Fan, J. B., Gunderson, K. L., Bibikova, M., Yeakley, J. M., Chen, J., Wickham Garcia, E., Lebruska, L. L., Laurent, M., Shen, R., and Barker, D. (2006). Illumina universal bead arrays. *Methods Enzymol*, 410:57-73.
- Fujisawa, H., Eguchi, S., Ushijima, M., Miyata, S., Miki, Y., Muto, T., and Matsuura, M. (2004). Genotyping of single nucleotide polymorphism using model-based clustering. *Bioinformatics*, 20:718-726.
- Hardiman, G. (2004). Microarray platforms-comparisons and contrasts. *Pharmacogenomics*, 5:487-502.
- Hoogendoorn, B., Norton, N., Kirov, G., Williams, N., Hamshire, M. L., Spurlock, G., Austin, J., Stephens, M. K., Buckland, P. R., Owen, M. J., and O'Donovan, M. C. (2000). Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Hum Genet*, 107:488-493.
- Jawaid, A., Bader, J. S., Purcell, S., Cherny, S. S., and Sham, P. (2002). Optimal selection strategies for QTL mapping using pooled DNA samples. *Eur J Hum Genet*, 10:125-132.

- Jawaid, A. and Sham, P. (2009). Impact and quantification of the sources of error in DNA pooling designs. *Ann Hum Genet*, 73:118-124.
- Kerr, G., Ruskin, H. J., Crane, M., and Doolan, P. (2008). Techniques for clustering gene expression data. *Computers in Biology and Medicine*, 38:283-293.
- Kirov, G., Nikolov, I., Georgieva, L., Moskvina, V., Owen, M. J., and O'Donovan, M. C. (2006). Pooled DNA genotyping on Affymetrix SNP genotyping arrays. *BMC Genomics*, 7:27.
- Kirov, G., Williams, N., Sham, P., Craddock, N., and Owen, M. J. (2000). Pooled genotyping of microsatellite markers in parent-offspring trios. *Genome Res*, 10:105-115.
- Knight, J., Saccone, S. F., Zhang, Z., Ballinger, D. G., and Rice, J. P. (2009). A comparison of association statistics between pooled and individual genotypes. *Hum Hered*, 67:219-225.
- Le Hellard, S., Ballereau, S. J., Visscher, P. M., Torrance, H. S., Pinson, J., Morris, S. W., Thomson, M. L., Semple, C. A., Muir, W. J., Blackwood, D. H., Porteous, D. J., and Evans, K. L. (2002). SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic Acids Res*, 30:e74.
- Lin, Y., Tseng, G. C., Cheong, S. Y., Bean, L. J., Sherman, S. L., and Feingold, E. (2008). Smarter clustering methods for SNP genotype calling. *Bioinformatics*, 24:2665-2671.
- Maouche, S., Poirier, O., Godefroy, T., Olaso, R., Gut, I., Collet, J.-P., Montalescot, G., and Cambien, F. (2008). Performance comparison of two microarray platforms to assess differential gene expression in human monocyte and macrophage cells. *BMC Genomics*, 9:302.
- Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the theory of statistics*, pp 181. McGraw-Hill, New York.
- Nakabayashi, K., Komaki, G., Tajima, A., Ando, T., Ishikawa, M., Nomoto, J., Hata, K., Oka, A., Inoko, H., Sasazuki, T., and Shirasawa, S. (2009). Identification of novel candidate loci for anorexia nervosa at 1q41 and 11q22 in Japanese by a genome-wide association analysis with microsatellite markers. *J Hum Genet*, 54:531-537.
- Norton, N., Williams, N. M., Williams, H. J., Spurlock, G., Kirov, G., Morris, D. W., Hoogendoorn, B., Owen, M. J., and O'Donovan, M. C. (2002). Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. *Hum Genet*, 110:471-478.
- Pearson, J. V., Huentelman, M. J., Halperin, R. F., Tembe, W. D., Melquist, S., Homer, N., Brun, M., Szelinger, S., Coon, K. D., Zismann, V. L., Webster, J. A., Beach, T., Sando, S. B., Aasly, J. O., Heun, R., Jessen, F., Kolsch, H., Tsolaki, M., Daniilidou, M., Reiman, E. M., Papassotiropoulos, A., Hutton, M. L., Stephan, D. A., and Craig, D. W. (2007). Identification of the genetic basis for complex disorders by use of pooling-based genomewide single-nucleotide-polymorphism association studies. *Am J Hum Genet*, 80:126-139.
- Perkel, J. (2008). SNP genotyping: six technologies that keyed a revolution. *Nat Meth*, 5:447-453.
- Risch, N. and Teng, J. (1998). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res*, 8:1273-1288.
- Sabatti, C. and Lange, K. (2005). Bayesian Gaussian Mixture Models for High Density Genotyping Arrays. *UC Los Angeles: Department of Statistics, UCLA*.
- Sham, P., Bader, J. S., Craig, I., O'Donovan, M., and Owen, M. (2002). DNA Pooling: a tool for large-scale association studies. *Nat Rev Genet*, 3:862-871.

- Simpson, C. L., Knight, J., Butcher, L. M., Hansen, V. K., Meaburn, E., Schalkwyk, L. C., Craig, I. W., Powell, J. F., Sham, P. C., and Al-Chalabi, A. (2005). A central resource for accurate allele frequency estimation from pooled DNA genotyped on DNA microarrays. *Nucleic Acids Res*, 33:e25.
- Tabeta, K., Shimada, Y., Tai, H., Ishihara, Y., Noguchi, T., Soga, Y., Takashiba, S., Suzuki, G., Kobayashi, T., Oka, A., Kobayashi, T., Yamazaki, K., Inoko, H., and Yoshie, H. (2009). Assessment of Chromosome 19 for Genetic Association in Severe Chronic Periodontitis. *Journal of Periodontology*, 80:663-671.
- Vens, M., Schillert, A., Konig, I., and Ziegler, A. (2009). Look who is calling: a comparison of genotype calling algorithms. *BMC Proceedings*, 3:S59.
- Visscher, P. M. and Le Hellard, S. (2003). Simple method to analyze SNP-based association studies using DNA pools. *Genet Epidemiol*, 24:291-296.
- Yau, C. and Holmes, C. C. (2008). CNV discovery using SNP genotyping arrays. *Cytogenetic and Genome Research*, 123:307-312.
- Zou, G. and Zhao, H. (2004). The impacts of errors in individual genotyping and DNA pooling on association studies. *Genet Epidemiol*, 26:1-10.
- Zuo, Y., Zou, G., and Zhao, H. (2006). Two-stage designs in case-control association analysis. *Genetics*, 173:1747-1760.